Influence des annotations imparfaites sur les systèmes de Traitement Automatique des Langues, un cadre applicatif : la résolution de l'anaphore pronominale

THÈSE

présentée et soutenue publiquement le 20/11/2008

pour l'obtention du

Doctorat de l'université Paris-Nord – Paris 13

(spécialité informatique)

par

Davy Weissenbacher

Composition du jury

Directeur :	Christophe Fouqueré
CoDirectrice :	Adeline Nazarenko
Examinateurs :	Jean-Gabriel Ganascia Sophia Ananiadou
Rapporteurs :	Bernard Victorri Philippe Leray

Laboratoire d'Informatique de Paris-Nord — CNRS UMR 7030

Remerciements

Je remercie Messieurs Bernard Victorri et Philippe Leray d'avoir accepté de rapporter cette thèse.

Je remercie Sophia Anianadou ainsi que Jean-Gabriel Ganascia pour avoir accepté d'examiner cette thèse.

Je remercie Christophe Fouqueré d'avoir dirigé mon travail. Ses remarques et son savoir ont considérablement enrichi mes reflexions et leurs perspectives, notamment en logique.

Je remercie chaleureusement Adeline Nazarenko d'avoir encadré mes travaux avec autant d'attention, de disponibilité et de bienveillance. Notre collaboration fut un plaisir, tant au plan professionnel que personnel.

Je remercie Henry Soldano qui s'est intéressé à mon travail dès le début et dont les conseils en matière d'apprentisage formel furent précieux.

Je remercie Daniel Kayser qui a répondu avec patience et clareté aux difficultés que j'ai rencontrées en logique.

Je remercie Thierry Poibeau qui m'a donné l'idée générale de cette thèse et pour ses leçons de linguistiques, dispensées lors de nos nombreuses mais fructueuses pauses café.

Je remercie Thierry Hamon pour son enseignement et son assistance quotidienne, même si le prix demandé fut une convertion au système Linux...

Je remercie Farid Nouioua pour ses remarques qui m'ont souvent donné à réfléchir.

Je remercie Sophie Aubin et Julien Derivière, qui ont facilité mon travail en m'apportant, toujours avec le sourire, leur aide et leur savoir-faire.

Enfin, je remercie ma mère, mon frère et mes amis pour leur présence et leur soutien, eux qui savent toujours me distraire et me divertir dès que je commence à travailler.

À ma Mère qui a attendu cette thèse avec impatience.

Table des matières

Liste des tableaux

Table des figures

Résumé

Un système informatique ne peut traiter automatiquement un texte sans qu'un certain nombre d'informations contenues dans ce texte, comme les mots ou les phrases, ne soient annotées. L'annotation est en général produite automatiquement par un processus incrémental. Pour ajouter un nouveau niveau d'annotation un système exploite des annotations d'entrée, plus "simples", produites par les systèmes qui ont été appliqués précédemment au texte. Or aucun système d'annotation, à ce jour, ne réalise automatiquement une annotation parfaite d'un texte. En raison de la complexité et de la diversité de la langue naturelle, le biais de représentation est élevé : il est difficile d'isoler les annotations d'entrée discriminantes qui sont nécessaires pour annoter un texte. De plus, ces annotations d'entrée, résultant elles-même d'un calcul automatique, sont imparfaites *i.e.* imprécises, bruitées et parfois manquantes. Enfin, le corpus sur lequel ont été choisies les annotations d'entrée discriminantes est un rassemblement artificiel de textes d'un même domaine qui donne une vue partielle de la langue. Cette erreur d'estimation implique une variation importante de la fiabilité de ces annotations lorsque le système est appliqué sur un corpus d'un autre domaine.

L'imperfection des annotations d'entrée et de sortie des systèmes de traitement automatique des langues (TAL) semble être une fatalité avec laquelle nous devons composer plutôt qu'un désagrément passager dû à la maturité naissante des techniques de TAL. Nous ne pouvons donc pas l'ignorer. L'imprécision des annotations se transmet et augmente à chaque niveau d'annotation du traitement d'un document. Si nous pouvons chercher à atténuer l'imperfection des informations d'entrée en réduisant le biais de représentation et l'erreur d'estimation, elles ne seront certainement jamais, sans intervention humaine, d'une qualité parfaite. Ce constat fait, une série de questions s'impose. Quelles sont les caractéristiques que le modèle de représentation des informations d'entrée doit réunir pour réduire le biais de représentation ? Comment adapter le statut des informations d'entrée dans le processus d'inférence du système lorsque, le système étant appliqué à un corpus d'un autre domaine, leur utilité vient à changer ? L'étude de la logique nous montre que nous savons raisonner formellement à partir d'informations imparfaites mais pouvons-nous en espérer un bénéfice pour le TAL ? Autrement dit, lequel de ces deux systèmes de TAL obtient les meilleures performances : un système qui intègre l'imperfection des informations d'entrée dans son processus de raisonnement ou un système prévu pour raisonner à partir d'informations d'entrée parfaites mais travaillant avec des informations imparfaites ?

Pour répondre à ces questions, nous avons proposé un classifieur reposant sur le modèle des réseaux bayésiens. Ce modèle d'inférence probabiliste est adapté pour raisonner sur les données de la langue naturelle. La possibilité de représenter dans un formalisme unique les attributs hétérogènes qui décrivent les annotations d'entrée diminue le biais de représentation. Tous les attributs pertinents pour une tâche d'annotation peuvent être exploités lors de l'inférence et

les données de chaque classe discriminées au mieux. Les probabilités conditionnelles *a priori* associées au réseau expriment un ensemble de contraintes dont nous nous servons pour estimer les valeurs inconnues des attributs et renforcer la fiabilité des attributs les plus bruités. De plus, leur apprentissage automatique atténue l'erreur d'estimation. L'étape d'apprentissage adapte les probabilités conditionnelles aux corpus et garantit que les attributs engagés dans la décision du classifieur sont toujours discriminants.

Nous avons validé notre modèle sur le problème de la résolution du pronom *it* anaphorique dans les textes anglais. Nous avons conçu et implémenté un classifieur pour la distinction des pronoms impersonnels des pronoms anaphoriques et pour le choix de l'antécédent puis, nous avons évalué les deux classifieurs sur des corpus de domaines et de genres différents.

L'analyse des résultats du classifieur pour la distinction des pronoms impersonnels a mis clairement en évidence l'intérêt de notre approche. Les résultats de notre classifieur sont meilleurs que ceux des systèmes de l'état de l'art concurrents.

Sur la tâche plus difficile du choix de l'antécédent, les résultats de notre second classifieur sont moins satisfaisants mais ils restent comparables aux résultats des systèmes de l'état de l'art : lorsque l'imperfection des attributs est prise en compte dans l'inférence, l'emploi d'attributs de mauvaise qualité ne dégrade que très faiblement les performances du système ; si, sans être parfaits, les attributs sont de meilleure qualité, leur utilisation améliore significativement les performances du système. De plus, l'analyse des erreurs du classifieur montre qu'il est encore trop simple et qu'il modélise imparfaitement le problème du choix de l'antécédent.

Pour corriger ces défauts du classifieur, nous proposons, en dernière partie de cette thèse, le prototype d'un nouveau classifieur reposant sur une extension des réseaux bayésiens dynamiques.

Chapitre 1

Introduction

Sommaire

1.1 Du document brut au document annoté

1.1.1 Définition de l'annotation

A la lecture d'un texte, l'homme, qui en comprend le sens, décode mentalement l'ensemble des informations contenues dans le texte. Il n'en est pas de même pour un système automatique. Pour permettre le traitement automatique, un certain nombre d'informations contenues dans le texte doivent être rendues explicites au système. Ces informations peuvent être très simples comme la distinction des mots et des phrases, ou bien plus complexes comme la structure grammaticale des phrases ou encore les relations de synonymie entre les syntagmes nominaux.

La mention explicite de ces informations exploitables par le système prend souvent la forme d'un ensemble d'annotations. Annoter une partie d'un document consiste à consigner un ensemble de propriétés, dont le type est prédéfinies, afin de caractériser une séquence continue ou discontinue du document. Ces propriétés peuvent être d'ordres différents comme des propriétés lexicales, syntaxiques, sémantiques voire ontologiques. On parle plus généralement d'annotation d'une partie d'un document, mais on peut aussi annoter le document lui même ou encore

le corpus. Notre définition s'étend sans difficulté. Annoter un document ou un corpus consiste à consigner un ensemble de propriétés prédéfinies afin de caractériser une information relative au document ou au corpus. La figure 1.1 donne un exemple d'un petit document brut (partie de gauche) qui est une simple séquence de caractères puis, dans la partie de droite, le même document où l'auteur, les mots et les phrases sont annotés avec une structure *xml*.

Texte Brut	Texte Annoté
Cyndi tasted the soup. She liked it.	<Document> <Nom>UnDocument</Nom> <Auteur>Grishman</Auteur> <Texte> <Phrase> <Mot>Cyndi</Mot> <Mot> tasted </Mot> <Mot>the </Mot> <Mot>soup</Mot> <Ponctuation>.</Ponctuation> </Phrase> <Phrase> <Mot>She</Mot> <Mot>liked</Mot> <Mot>it</Mot> <Ponctuation>.</Ponctuation> </Phrase> <Texte> </Document>

FIG. 1.1 – Un texte brut puis annoté avec un format *xml*

La richesse de l'annotation d'un document est variable et dépend de l'utilisation qui doit en être faite. Il s'agit souvent d'un compromis entre la qualité recherchée des réponses du système, et la rapidité de la réponse attendue. Un système de recherche d'information rudimentaire qui doit traiter de gros volumes de données rapidement pourrait se satisfaire d'un corpus où seules les entités nommées (EN) sont annotées. Les EN sont de bons indicateurs du sujet général d'un document. Le nom de *Nicolas Sarkozy* retrouvé dans un document suffit à le distinguer d'un document ou le nom *Linux* apparaîtrait. Pour un système d'extraction d'information, cette annotation serait insuffisante et demanderait à être complétée par d'autres annotations, comme la recherche des hyperonymes des entités nommées ou de classes sémantiques de verbes spécifiques. Pour cette tâche, l'utilisateur tolère un délai de réponse plus important mais exige des réponses précises et justes.

1.1.2 Représentation d'une annotation

Le schéma d'annotation

Un schéma d'annotation est une description abstraite de la propriété que l'on souhaite annoter. Elle précise le nom général de l'annotation, les attributs devant décrire l'élément annoté et peut aussi spécifier des contraintes sur les valeurs possibles des attributs. Chaque annotation est

une réalisation particulière et conforme du schéma d'annotation qui la définit. Il n'y a aucune limitation au nombre d'annotations que l'on peut créer, sinon celles imposées par le document (par exemple si nous ne désirons annoter que les mots d'un document nous allons créer une annotation pour chaque mot du document, ni plus ni moins).

Annotation intégrée ou déportée

Les annotations peuvent être intégrées au document brut comme dans le document de la figure 1.1 ou, au contraire, être déportées dans un fichier ou une zone séparés comme pour le document du tableau 1.1. Dans le cas d'une annotation déportée, un adressage des séquences du document brut les identifie dans le document annoté. Chaque type d'annotation, intégrée ou déportée, présente des avantages et des inconvénients [Nazarenko *et al.*, 2006]. Nous avons choisi pour ce travail de déporter les annotations. Deux raisons principales motivent notre choix. Nous verrons dans la section 2.2.2 que le nombre et le niveau des annotations nécessaires pour réaliser notre tâche sont encore imparfaitement connus. La gestion déportée des annotations autorise l'ajout de nouveaux niveaux d'annotations sans modification des niveaux existants. Nous verrons aussi dans la section 1.2.3 que la gestion d'annotations concurrentes est un problème qui ne peut être ignoré lors de la résolution d'une tâche de TAL. En déportant les annotations produites par différents outils de TAL concurrents nous pouvons représenter l'intégralité des annotations qu'elles soient cohérentes entre elles ou non.

Un langage de description pour le schéma d'annotation

Un schéma d'annotation peut être exprimé par différentes grammaires. Si les langages à balisages du type *DTD* ou *XML schema* sont couramment employés, ils ne sont pas les plus expressifs. Les grandes plateformes d'annotations, telle que GATE ou encore UIMA, lui ont préféré pour cette raison le langage de représentation objet [Thayse *et al.*, 1990a]. Nous présentons rapidement les points principaux du document technique TIPSTER de [Grishman, 1997] sur lequel les concepteurs de ces plateformes se sont appuyés.

Dans ce langage, un schéma d'annotation est décrit par la classe d'objet *Annotation*. Chaque annotation est une instance particulière de cette classe. L'annotation est identifiée par un numéro d'identification unique et d'un certain type. Elle est ancrée dans le document en précisant le début et la fin de la séquence qu'elle annote (si la séquence peut être discontinue, par exemple, pour annoter des verbes à particules, l'ancre précise les bornes de chaque segment de la séquence). Chaque annotation contient un ensemble d'attributs qui précise les informations apportées par l'annotation. Ces attributs peuvent avoir pour valeur une chaîne particulière de caractères ou un ensemble de références sur d'autres annotations.

Les annotations relatives à un document du corpus sont décrites par la classe *Document* et regroupées dans un objet de cette classe propre au document. L'objet est aussi identifié univoquement et rattaché au corpus auquel le document appartient. Un ensemble d'annotations spécifie certaines informations propres au document (*ex.* l'auteur du document, sa date de création...).

Enfin le corpus est un objet de la classe *Collection*. Il comprend les références des documents qui le composent et un ensemble d'annotations pour décrire les informations qui ne sont

pas localisées au niveau du document mais au niveau du corpus, des informations statistiques telles que le nombre d'occurences d'un bigramme par exemple.

Nous reproduisons ici l'exemple de [Grishman, 1997] montrant une phrase où les mots[1], les entités nommées et la syntaxe sont annotés. Les objets sont représentés sous la forme compacte d'un tableau. Les constituants de la phrase et de l'analyse syntaxique sont désignés par une liste des numéros d'identification des annotations.

texte					
Cyndi tasted the soup.					
Annotation					
Mot_1	ID :1	TYPE :Mot	AncreDébut :0	AncreFin :5	POS :NP
Mot_2	ID :2	TYPE :Mot	AncreDébut :7	AncreFin :12	POS :VBD
Mot_3	ID :3	TYPE :Mot	AncreDébut :14	AncreFin :16	POS :DT
Mot_4	ID :4	TYPE :Mot	AncreDébut :18	AncreFin :21	POS :NN
EN_5	ID :5	TYPE :Entité Nommée	AncreDébut :0	AncreFin :5	Classe :Personne
Phrase_6	ID :6	TYPE :Phrase	AncreDébut :0	AncreFin :22	Constituant :{[1],[2],[3],[4]}
Analyse_7	ID :7	TYPE :Elément syntaxique	AncreDébut :0	AncreFin :5	Etiquette : NP Constituant :{[1]}
Analyse_8	ID :8	TYPE :Elément syntaxique	AncreDébut :14	AncreFin :21	Etiquette : NP Constituant :{[3],[4]}
Analyse_9	ID :9	TYPE :Elément syntaxique	AncreDébut :7	AncreFin :12	Etiquette : NP Constituant :{[2]}
Analyse_10	ID :10	TYPE :Elément syntaxique	AncreDébut :7	AncreFin :21	Etiquette : VP Constituant :{[2],[9]}
Analyse_11	ID :11	TYPE :Elément syntaxique	AncreDébut :0	AncreFin :22	Etiquette : S Constituant :{[8],[10]}

TAB. 1.1 – Exemple d'un texte annoté avec la représentation objet

1.2 L'annotation, un processus incrémental imparfait

1.2.1 Le processus d'inférence pour l'annotation

Les tâches de TAL consistent en majorité à ajouter une nouvelle catégorie d'annotations au document ou au corpus. A l'exception des tâches d'annotation les plus "simples" comme la segmentation en mots ou en phrases qui peuvent être effectuées à partir du document brut, les systèmes doivent s'appuyer sur un ensemble d'annotations d'entrée pour calculer et apposer leur propres annotations. Ces annotations produites seront, ensuite, utilisées à leur tour comme annotations d'entrée pour d'autres tâches. Les valeurs des attributs des annotations d'entrée servent de connaissances au processus d'inférence du système dont la décision détermine l'ajout et les valeurs d'une nouvelle annotation ou, au contraire l'inaction.

[1]Les étiquettes associées aux mots appartiennent au jeu d'étiquettes défini pour le projet Tree Bank (Université de Pennsylvania)

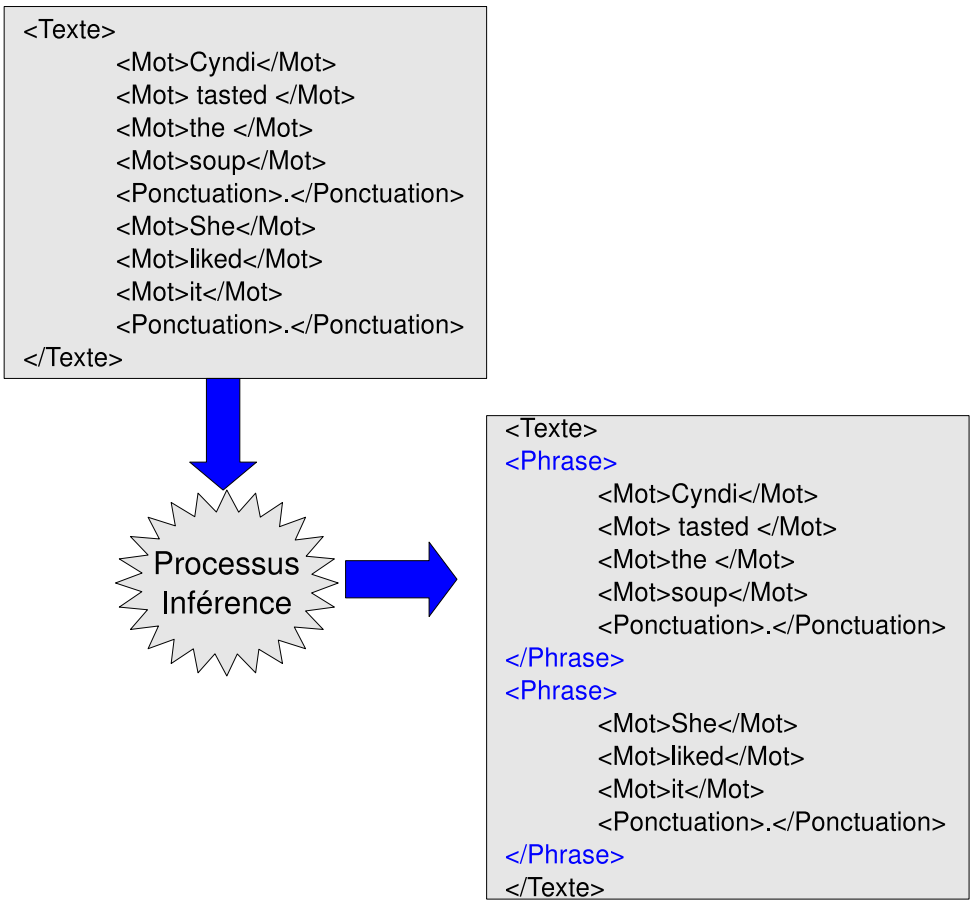

FIG. 1.2 – Un processus d'inférence pour la segmentation en phrases

Supposons un système dont la tâche est de segmenter un document en phrases. Les annotations d'entrée de notre système sont les annotations marquant le début et la fin du document ainsi que l'ensemble des mots. Nous représentons les annotations sous le format XML pour simplifier l'exemple. Le processus d'inférence de notre système se réduit à l'application de deux règles logiques sur l'ensemble des séquences continues correspondant aux mots du document[2] :

- Règle 1 : Si la séquence commence au début du document et finit par un mot suivi d'un point alors la séquence est une phrase.
- Règle 2 : Si la séquence commence après un point et finit par un mot suivi d'un point alors la séquence est une phrase.

L'application de notre système sur un document est schématisée par la figure 1.2

[2]Nous faisons l'économie d'une présentation formelle du processus d'inférence par souci de clarté de notre exposé. Toutefois notre tâche se reformule très bien par une théorie logique du premier ordre et le processus d'inférence par le système de dérivation classique.

1.2.2 L'imperfection des annotations d'entrée

Le processus d'inférence de notre système jouet segmente correctement notre document en insérant les balises de début et de fin de phrase aux positions attendues dans le document. Mais nous avons tacitement supposé que les annotations d'entrée et les règles du processus d'inférence sont, non seulement correctes, mais aussi suffisantes. Deux conditions qui sont rarement réunies lorsque nous appliquons un système de TAL sur un document. Les sections suivantes tentent d'en éclaircir les raisons en présentant le problème comme un problème d'apprentissage [Cornuéjols & Miclet, 2002].

Corpus d'acquisition et règles d'inférences

La plupart des tâches d'annotation consistent en fait à étiqueter des entités textuelles (mots, syntagmes, phrases, paragraphes, *etc.*) et se ramènent ainsi à des problèmes de classification. On distingue généralement trois étapes dans la conception d'un système de classification. Dans la première étape on se dote d'un ou de plusieurs corpus d'acquisition. Les séquences de ces corpus à annoter, que l'on appelle aussi exemples d'entraînement, apportent un ensemble d'observations qui permettent au concepteur d'isoler les attributs pertinents pour la tâche d'annotation qu'il s'est fixée. Dans la deuxième étape, ayant adopté un ensemble d'attributs, le concepteur formule les règles d'inférences pour distinguer les séquences à annoter des autres séquences. Si le concepteur dispose de corpus préalablement étiquetés, il peut apprendre automatiquement ces règles d'inférences. La dernière étape est une étape d'évaluation pour laquelle le concepteur doit se munir d'un nouveau corpus distinct des corpus d'acquisition, le corpus de test. Le concepteur applique ses règles d'inférence pour assigner une classe aux nouvelles séquences de ce corpus. Pour que les règles puissent être appliquées à ces nouvelles séquences celles-ci doivent être décrites avec les mêmes attributs que ceux décrivant les exemples d'entraînement. La performance du système se calcule alors sur le nombre de séquences correctement classées.

Un biais de représentation élevé

Le schéma d'annotation pertinent pour la formulation des règles qui nous permettraient de distinguer parfaitement les séquences à annoter des autres séquences est souvent inconnu et difficile à définir. Lorsque nous adoptons un schéma au détriment d'un autre nous introduisons un biais de représentation. Le corpus d'acquisition sur lequel porte l'analyse conditionne en partie la sélection des annotations et de leurs attributs, alors qu'il ne donne jamais qu'une vue partielle de la langue qu'on cherche à modéliser. Le choix des annotations repose sur les seuls exemples rencontrés et attestés dans le corpus d'acquisition qui ne constitue pas un ensemble exhaustif. Il n'est donc pas certain que les annotations choisies puissent réellement discriminer toutes les séquences que nous souhaitons annoter dans le corpus de test.

Analysons le comportement du système jouet sur l'exemple suivant tiré d'un résumé d'un article de génomique :

> *Haemolysin BL (HBL) is composed of a binding component, B, L1 and L2. The first gene of the hbl operon, hblC, in the B. Cereus type strain, ATCC 14579, was inactivated in this study.*

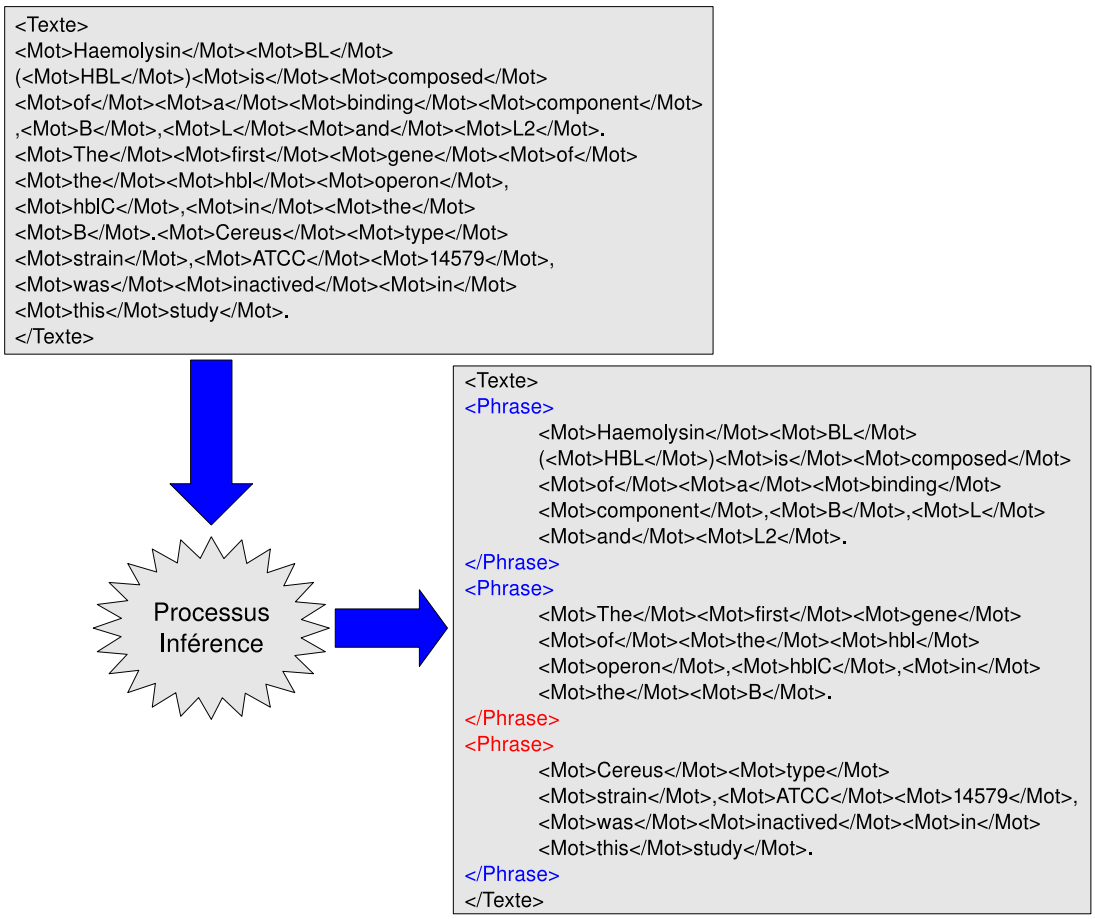

FIG. 1.3 – Echec du processus d'inférence pour la segmentation en phrases

Le système doit segmenter la séquence en raisonnant avec les mêmes règles. Il segmente correctement la première mais pas la deuxième. En effet l'annotation lui indique, à juste titre, que la séquence 'B' est un mot comme le *B* de la séquence *a binding component, B, L1 and...*. En utilisant cette information il segmente la séquence qui contient l'abréviation *B[acillus]. Cereus* (voir la figure 1.3).

Pour éviter cette erreur il faut annoter les abréviations et ajouter de nouvelles contraintes sur ces attributs aux règles d'inférences du système[3]. Les règles pourraient se reformuler ainsi :

– Règle 1 : Si la séquence commence au début du document et se termine par un mot *différent d'une abréviation* suivi d'un point alors la séquence est une phrase.
– Règle 2 : Si la séquence commence après un point et finit par un mot *différent d'une abréviation* suivi d'un point alors la séquence est une phrase.

[3]Notons que la précision de l'attribut *Commencer par une majuscule* dans l'annotation des mots et l'ajout de cette contrainte à la règle de segmentation n'aurait pas corrigé le problème à elle seule. Le mot *Cereus* commence par une majuscule.

La fiabilité des annotations

Une fois le schéma d'annotation défini pour réaliser une tâche d'annotation, rien ne garantit que nous soyons capable de calculer les annotations de ce nouveau niveau avec une grande fiabilité.

La fiabilité d'une annotation concerne à la fois l'identification et la délimitation du segment à annoter. L'erreur d'identification est une séquence du document qui aurait du être annotée ou ne pas l'être et qui a été ignorée ou, au contraire, ajoutée. Un système de reconnaissance des entités nommées peut ignorer certaines entités du texte et annoter certains mots qui ne le sont pas. L'erreur dans la délimitation de la séquence est plus simple. Un système de segmentation des phrases d'un document peut, comme nous l'avons vu, confondre un point d'abréviation avec un point de fin de phrase, et découper incorrectement la phrase (ce qui entraîne, dans le cas de la segmentation en phrases, la création d'une nouvelle phrase incorrecte).

La fiabilité des attributs est aussi ambivalente avec l'impossibilité d'attribuer une valeur à un attribut ou l'attribution d'une valeur erronée. Par exemple, lors d'une tâche d'analyse morpho-syntaxique l'analyse de certains mots échoue et l'attribut marquant leur classe grammaticale reste incomplet. Et parmi les mots analysés, la valeur reçue par l'attribut peut être fausse (*e.g.* la valeur *adjectif* pour le mot *bleu* alors qu'il est employé comme substantif comme dans la phrase *J'ai un bleu.*).

L'erreur d'estimation

Enfin, même si les annotations et leurs attributs sont pertinents et fiables pour la tâche, l'erreur d'estimation peut faire chuter les performances du système. Le corpus sur lequel l'expert effectue son analyse n'est pas un échantillon de données tiré aléatoirement. Le corpus est le plus souvent une construction artificielle [Habert *et al.*, 1997] composée de textes d'un même domaine (par ex. la génomique, des dépêches économiques...) rassemblés en fonction d'objectifs applicatifs ou en fonction de la disponibilité des textes. Or le rôle d'une annotation ou d'un attribut peut varier suivant le domaine du corpus. Un attribut qui se révèle être un très bon indicateur pour annoter un corpus d'un domaine peut être moins bon ou même erroné sur un corpus d'un autre domaine. Dans un roman, hormis les noms propres, peu de mots commencent par une majuscule. Lorsqu'un mot du dictionnaire des noms communs commence par une majuscule et est rencontré après un point suivi d'un espace, ce mot débute une phrase avec une grande probabilité. Cette probabilité est plus faible en présence des références bibliographiques, des abréviations, des mesures physiques ou encore des noms de gènes qui sont ambigus avec des mots de la langue courante (et dont la nomenclature impose de commencer par une majuscule) d'un article de génomique. Nous rencontrerons à de multiples reprises des variations de ce type dans nos expériences.

1.2.3 Comment structurer des annotations imparfaites ?

Les sections précédentes mettent en lumière la difficulté de travailler à partir d'annotations parfaites et de produire des annotations de cette qualité. En l'absence de correction humaine, la nécessité d'exploiter des annotations imparfaites s'impose. Il est donc important d'intégrer cette information au sein de la représentation des annotations.

La représentation que nous avons décrite dans la section 1.1.2 ne permet pas de représenter la fiabilité des attributs et des annotations. Nous devons modifier le modèle afin d'exprimer cette nouvelle information.

Pour tenir compte des attributs manquants il suffit de marquer leur absence par une valeur spéciale, la valeur "*Unknown*", une option qui est déjà proposée par certains outils d'annotation.

Les annotations manquantes ou inattendues posent plus de difficultés. Une stratégie naturelle, qui a été retenue pour gérer l'ambiguïté des annotations morpho-syntaxiques [sub-committee TC37 SC4, 2005 08 22], consiste à annoter le document avec plusieurs annotations concurrentes (provenant d'un même outil ou d'outils concurrents). La représentation doit alors accepter la multi-annotation et être enrichie pour exprimer le degré de confiance que le système accorde aux valeurs prises par les attributs des annotations concurrentes.

Pour exprimer la multi-annotation nous ajoutons une classe d'objets supplémentaire. Cette classe ne compte qu'un unique attribut *constituant*. Cet attribut contient une liste de références sur l'ensemble des annotations qui sont en concurrence. Pour spécifier le degré de confiance du système dans les valeurs de chaque attribut incertain, nous ajoutons un attribut contenant la valeur de cette confiance pour chacun des attributs incertains (voir le tableau 1.2. La nature de la valeur de confiance fera l'objet de notre étude dans le chapitre 3.

L'exemple qui suit montre comment représenter des analyses concurrentes qui manifestent l'incertitude de notre système sur les valeurs du rôle morpho-syntaxique et de la classe de l'entité nommée d'un mot donné. Nous considérons artificiellement dans cet exemple que les valeurs des attributs sont indépendantes entre elles, ce qui est faux. Si nous arrêtons notre choix sur la valeur *Société* alors la valeur de l'attribut POS est certainement un nom propre (*NP*) et la valeur du nom commun (*NN*) devient nettement improbable. Nous reviendrons longuement sur ce phénomène dans le chapitre 3

Phrase : `Le [Club] a augmenté ses prix.`	
Annotation_Concurrente	
Mot_2_Concurrence	Constituant :[3],[4],[5],[6]
	Annotation
Constituant_3	POS :NP, EN_Classe :Société, Croyance_POS :70%, Croyance_EN_Classe :60%
Constituant_4	POS :NN, EN_Classe :Aucune, Croyance_POS :70%, Croyance_EN_Classe :40%
Constituant_5	POS :NP, EN_Classe :Aucune, Croyance_POS :30%, Croyance_EN_Classe :40%
Constituant_6	POS :NN, EN_Classe :Société, Croyance_POS :30%, Croyance_EN_Classe :60%

TAB. 1.2 – Exemple d'une annotation concurrente

1.3 Problématique et contexte de la thèse

1.3.1 Problématique de thèse : un système de résolution des anaphores reposant sur des annotations imparfaites

L'annotation est, en général, vue comme un processus incrémental. Les systèmes d'annotations sont appliqués aux documents séquentiellement. Pour ajouter son niveau d'annotation

un système a recours aux annotations, plus “simples”, produites par les systèmes qui ont été appliqués précédemment au document.

Or, aucun système d’annotation, à ce jour, ne réalise automatiquement une annotation parfaite sur tous les types de corpus. La difficulté de modéliser les connaissances d’un système sans biais de représentation, la fiabilité, toujours suspecte, des annotations sur lesquelles le système repose ou encore l’erreur d’estimation sensible à chaque emploi du système sur un corpus d’un domaine nouveau, détériorent la qualité des annotations qu’il produit. Les erreurs d’annotation se propagent et en provoquent de nouvelles à chaque niveau d’annotation du traitement d’un document.

L’imperfection des annotations d’entrée et de sortie des systèmes de TAL semble être une fatalité avec laquelle nous devons composer plutôt qu’un désagrément passager dû à la maturité naissante des techniques de TAL que nous pouvons ignorer. Si nous pouvons chercher à atténuer l’imperfection des informations d’entrée en réduisant le biais de représentation et l’erreur d’estimation, elles ne seront certainement jamais, sans intervention humaine, d’une qualité parfaites. Ce constat fait, une série de questions s’impose.

Quelles sont les caractéristiques que le modèle de représentation des informations d’entrée doit réunir pour réduire le biais de représentation ? Comment adapter le statut des informations d’entrée dans le processus d’inférence du système lorsque, le système appliqué à un corpus d’un autre domaine, leur utilité vient à changer ? L’étude de la logique nous montre que nous savons raisonner formellement à partir d’informations imparfaites mais pouvons-nous en espérer un bénéfice pour le TAL ? Autrement dit, lequel de ces deux systèmes de TAL obtient les meilleures performances : un système qui intègre l’imperfection des informations d’entrée dans son processus de raisonnement ou un système prévu pour raisonner à partir d’information d’entrée parfaites mais travaillant avec des informations imparfaites ?

Il est impossible de présenter et de défendre une réponse générale à ces questions pendant un travail de thèse. Pour cette raison nous avons circonscrit notre examen à une tâche précise du TAL : la résolution d’anaphore. Cette tâche d’annotation est complexe et requiert pour sa résolution des annotations de natures et de difficultés de calcul variées. Des caractéristiques qui, nous le verrons dans la suite de notre travail, font de cette tâche un très bon sujet d’étude.

Nous avons organisé notre thèse de la manière suivante :

Dans le chapitre suivant nous définissons et exemplifions le phénomène linguistique de l’anaphore. Nous présentons les approches de résolution des anaphores pronominales les plus connues de l’état de l’art et une liste non exhaustive des annotations qu’elles emploient pour y parvenir.

Le chapitre 3 clarifie la notion d’annotation imparfaite. Une annotation est imparfaite parcequ’elle est imprécise, incertaine ou encore manquante. Si les annotations des systèmes de TAL peuvent être imprécises, cette imperfection reste marginale, à la différence des annotations incertaines ou manquantes qui sont courantes dans notre domaine. Nous simplifions notre problème en limitant notre recherche aux représentations d’informations incertaines ou absentes. Parmi les choix possibles nous motivons notre préférence pour une représentation numérique de l’incertitude par les probabilités.

De nombreux formalismes ont été proposés pour raisonner sur des informations probabilisées. Dans le chapitre 4 nous rappelons les formalismes les plus connus avant de nous déterminer pour le modèle des réseaux bayésiens. Ce modèle est resté peu utilisé dans la communauté du TAL bien qu’il apporte des réponses élégantes au raisonnement à partir d’informations in-

certaines et manquantes. Nous détaillons sur des exemples simples les différents mécanismes offerts par ce formalisme.

Avant de proposer un système de résolution complet, nous avons cherché à valider notre modèle sur un problème plus simple par le nombre et la complexité des annotations en jeux : la distinction des pronoms impersonnels des pronoms anaphoriques. Nous expliquons notre démarche et décrivons le classifieur que nous avons implémenté dans le chapitre 5. Les expériences concluantes que nous avons menées et qui sont décrites à la fin de ce chapitre nous ont encouragé à poursuivre l'implémentation d'un système de résolution complet.

Nous commentons nos choix de conception de ce système au début de l'avant dernier chapitre et évaluons ce dernier dans une série d'expériences dont nous discutons les résultats. Nous terminons notre travail en généralisant les conclusions que nous avons obtenues sur le problème particulier de la résolution des anaphores aux autres tâches d'annotation.

1.3.2 Contexte de la thèse : le projet ALVIS

La problématique de cette thèse est s'inscrit dans une réflexion plus générale à laquelle nous avons participé durant notre collaboration au projet ALVIS[4]. ALVIS est un projet européen dont l'objectif était de concevoir un moteur de recherche sémantique spécialisé reposant sur une architecture pair à pair. Du point de vue du TAL, l'enjeu de ce projet était de mesurer l'impact d'une analyse linguistique dans la recherche d'information spécialisée. Car, si nous savons que l'intégration des outils de TAL dans un moteur de recherche généraliste n'est pas forcément concluante, son intérêt pour un moteur de recherche spécialisé est encore mal connu.

Ce projet propose une architecture ouverte, où chaque pair est responsable de la collecte et de l'enrichissement des documents qu'il doit indexer avant de les partager avec le réseau. L'indépendance des pairs permet de les spécialiser pour un domaine particulier. Un pair dispose d'une plateforme d'analyse linguistique que l'on peut enrichir selon les besoins particuliers induit par le domaine des documents que doit traiter le pair. La conception de cette plateforme linguistique a soulevé le problème de l'articulation de différents outils de TAL qui la composent [Nazarenko *et al.*, 2006]. Leurs formats d'entrée et de sortie diffèrent et il est important de pouvoir remplacer un outil au profit d'un autre outil mieux adapté sans avoir à modifier la plateforme elle-même. La solution apportée par le projet a été l'encapsulation de chaque outil dans un module devant normaliser les formats d'entrée et de sortie de l'outil selon un format d'annotation XML général défini pour ce projet.

C'est dans le cadre de cette réflexion sur l'articulation des outils de TAL que nous avons énoncé le problème de l'imperfection des annotations produites par ces outils. Pour des raisons de simplicité et de rapidité de calcul, l'intégration de l'imperfection des annotations à été ignorée dans la conception de la plateforme mais il nous a paru important de poursuivre en parallèle cette recherche afin d'apporter les premiers résultats nécessaires à une évolution possible de la plateforme.

En raison de notre participation à ce projet nous avons du travailler sous certaines contraintes. Nous avions l'obligation d'intégrer notre système de résolution des anaphores (que nous présentons dans le chapitre 6) dans la plateforme linguistique d'un pair spécialisé pour le traitement de documents de génomique. Nous avons donc choisi de travailler avec les outils développés

[4]Référence du projet : STREP IST-1-002068, 2004-2006, `http://www.alvis.info`.

pour cette plateforme, la plateforme *Ogmios* (que nous présentons dans la section 6.1.3), et de réaliser une partie de nos expériences sur un corpus de génomique (que nous décrivons dans la section 6.1.3). Nous avons ainsi bénéficié des outils produits pendant le projet et d'une coopération active de plusieurs chercheurs et ingénieurs de ce projet. Cette coopération s'est traduite par de nombreuses discussions qui ont orientées ce travail et par une aide concrète dans sa réalisation. Nous souhaitons remercier nommément Adeline Nazarenko, Sophie Aubin et Séverine Popek pour leurs annotations et leurs participations à la correction des données, Thierry Hamon, Thierry Poibeau, Julien Derivière et Eric Alphonse pour leurs aides dans le prétraitement des corpus nécessaires à nos expériences.

Chapitre 2

La résolution des anaphores pronominales

Sommaire

L'anaphore est un phénomène linguistique complexe et varié. Dans ce chapitre, nous étudions rapidement ce phénomène avant de justifier notre choix de travailler uniquement sur la résolution des relations anaphoriques du pronom *it*. Puis nous présentons les différentes approches ainsi que les différents systèmes de résolution des anaphores proposés par l'état de l'art, qui ont nourri notre réflexion durant ce travail.

2.1 L'anaphore

2.1.1 Définition

Le mot *anaphore* vient du grec ancien $\alpha\nu\alpha\phi o\rho\alpha$. Il est composé du préfixe $\alpha\nu\alpha$ qui signifie retour, en arrière, et de $\phi o\rho\alpha$ qui porte, présente. L'anaphore est une relation linguistique entre deux entités du discours, *l'entité anaphorique* (que l'on appelle *anaphore*) et *l'antécédent*. La relation est définie lorsque l'anaphore réfère à l'antécédent. Dans la phrase

[Le chien]$_1$ a mangé sa pâtée puis [il]$_1$ est allé se coucher

le pronom *il* est anaphorique et réfère au groupe nominal *Le chien*.

2.1.2 Les types d'anaphores

La diversité et les caractéristiques des anaphores rencontrées dans un document dépendent de la langue de ce dernier. Nous focaliserons notre étude sur la langue anglaise, la langue naturelle qui a été la plus étudiée jusqu'à présent dans le domaine du TAL.

Les variétés des anaphores dépendent de la nature et de la position de l'antécédent. Nous étudions les anaphores les plus communes dans cette section. Pour une étude détaillée des anaphores, le lecteur interressé pourra se reporter à l'étude de référence de [Huang, 2000].

Anaphore et Coréférence

Dans une relation de coréférence les entités textuelles désignent un même objet du monde réel ou d'un monde imaginaire, par exemple les entités de la coréférence suivante

After the first round [Dominique Voynet]$_1$ knows that [she]$_1$ will not win the election.

réfèrent toutes deux à la personne politique française.

L'anaphore et la coréférence sont liées comme le montre l'exemple précédent où la relation de coréférence est aussi anaphorique. Mais, selon [Mitkov, 2002], elles ne sont pas identiques, certaines relations peuvent être anaphoriques sans être coréférentes et inversement. Pour défendre sa position il compare deux contextes. Dans le contexte :

John has his own destiny.

nous pouvons substituer l'antécédent à l'anaphore sans modifier le sens de la proposition :

John has his own destiny.

Une telle substitution est impossible sans changer le sens de la proposition dans le contexte suivant :

Every man has his own destiny.

Pour trouver un exemple de relation de coréférence qui ne soit pas anaphorique, il suffit de considérer deux documents qui parlent d'une même personalité politique. Les entités textuelles se réfèrent bien au même objet du monde et ne sont pas anaphoriques puisqu'il s'agit de deux documents différents.

Nous discutons le statut de cette égalité uniquement dans le but de distinguer les deux relations. Ce problème est d'ordre linguistique et n'interresse qu'indirectement ce travail qui ne différenciera pas les deux relations lors de la résolution de la relation, voir 2.1.3.

Type d'anaphore déterminé par la nature de l'entité anaphorique

Un grand nombre d'entités peuvent introduire une relation anaphorique. La catégorie de la relation anaphorique est en partie déterminée par la nature syntaxique de l'anaphore. Nous décrivons les grandes catégories d'anaphores sans entrer dans les détails et les cas particuliers de ces phénomènes. Nous illustrons ces types d'anaphores par des exemples que nous avons extrait de deux corpus composés de textes techniques et des résumés d'articles de génomique (voir la section 6.1.3 pour une présentation détaillée de ces corpus).

L'anaphore la plus connue est l'anaphore pronominale. Tous les pronoms à la 3^{eme} personne, singuliers ou pluriels, peuvent être anaphoriques :

– Pronom personnel : *The aim of this document is to serve as an introduction to the technologies which are available to make [Linux]$_1$ usable by people who, through some disability would otherwise have problems with [it]$_1$.*

- Pronom réflexif : *The main disadvantages of [xzoom]$_1$ are that it can't magnify under [itself]$_1$, that some of the key controls aren't compatible with fvwm, the normal Linux window manager and that its default configuration doesn't run over a network.*
- Pronom possessif : *The purpose of The BLINUX Documentation and Development Project is to serve as a catalyst which will both spur and speed the development of software and documentation which will enable [the blind user]$_1$ to run [his]$_1$ or her own Linux workstation.*
- Pronom démonstratif : *We also found an orf, [orf16]$_1$, upstream of grpE in the archaeon Methanosarcina mazei S-6, but [this]$_1$ gene differs from the eubacterial counterpart.*
- Pronom relatif : *The aim of this document is to serve as an introduction to [the technologies]$_1$ [which]$_1$ are available to make Linux usable by people who, through some disability would otherwise have problems with it.*

L'anaphore nominale est une reprise anaphorique d'un syntagme nominal par un syntagme nominal défini. La reprise anaphorique peut apporter des informations supplémentaires sur l'antécédent si une relation sémantique unit l'anaphore et l'antécédent comme une relation de synonymie, d'hyperonymie, d'hyponymie, *etc*. Dans l'exemple suivant l'anaphore est un hyperonyme de l'antécédent :

> *[xocr]$_1$ is [a package]$_1$ which implements optical character recognition for Linux.*

Un verbe peut être anaphorique comme nous l'avons vu dans l'exemple de la section 2.1.2, un adverbe aussi comme le montre l'exemple suivant :

> *Will you walk with me to [the garden]$_1$? I've got to go down [there]$_1$ and Bugs has to go to the longhouse.*

La dernière grande catégorie d'anaphores est appelée *anaphore zéro* ou *ellipse*. Dans une ellipse l'anaphore est omise et le sens de la phrase ne peut être retrouvé que par l'ajout de l'antécédent à la position supposée de l'anaphore. Les formes les plus communes d'ellipses sont les ellipses pronominales, nominales et verbales. Nous donnons un exemple d'une ellipse nominale qui aurait pu tout aussi bien être pronominale :

> *There are several known screen magnification programs, xmag which will magnify a portion of the screen as much as needed but ∅ is very primitive.*

Type d'anaphore déterminé par la nature de l'antécédent

Un autre facteur intervient dans la catégorisation de l'anaphore, il s'agit de la nature de l'antécédent. Dans la majorité des relations anaphoriques l'antécédent est un syntagme nominal, mais il peut être d'une nature syntaxique différente :

- des antécédents coordonnés : *[The AbrB, Hpr and Sin proteins]$_1$ are the best-studied examples of these regulatory molecules. Their role is to prevent inappropriate and possibly detrimental functions from being expressed during exponential growth when [they]$_1$ are not needed.*
- une proposition : *Previously I said that [there had to be something better than xmag]$_1$, well [this]$_1$ is it.*
- une phrase ou un groupe de phrases : *[In principle it should be possible to put together a complete, usable Linux system for a visually impaired person for about $500]$_1$. I have yet to see [this]$_1$.*

[Amsili *et al.*, 2005] opposent les anaphores abstraites, les anaphores dont l'antécédent est une entité abstraite comme une phrase ou un groupe de phrases, et les anaphores individuelles, dont l'antécédent est une entité comme un syntagme nominal.

Type d'anaphore déterminé par la position de l'antécédent

La distance de l'antécédent par rapport à celle de l'anaphore peut varier selon le genre du corpus. Lorsque l'anaphore et l'antécédent sont situés dans la même phrase l'anaphore est dite intraphrastique et interphrastique lorsqu'ils ne partagent pas la même phrase.

Même si l'antécédent se trouve en grande majorité dans la même phrase ou dans la phrase précédente, il est toutefois impossible de limiter une fenêtre de recherche pour l'antécédent. [Mitkov, 2002] résume les statistiques obtenues par différents auteurs sur des corpus variés. La plus grande distance observée entre l'anaphore et l'antécédent est de 15 phrases.

Enfin la position de l'anaphore relativement à celle de l'antécédent n'est pas fixe. Lorsque l'anaphore précède l'antécédent comme dans l'exemple :

> *[She]$_1$ is now as famous as her ex-boyfriend. From the deserts of Kazakhstan to the south seas of Tonga, everyone knows [Monica Lewinshy]$_1$*[5]

la relation est appelée cataphore.

Les chaînes de coréférences

La relation de coréférence étant une relation d'équivalence, il est possible de partitionner les mots d'un texte en différentes classes de coréférences. Chaque classe réfère à un objet du discours distinct. Nous pouvons ainsi distinguer, en suivant [Mitkov, 2002] deux classes de référents, une première pour l'actrice Sophia Loren et une seconde pour le chanteur Bono :

> *[Sophia Loren]$_1$ says [she]$_1$ will always be grateful to [Bono]$_2$. [The actress]$_1$ revealed that [the U2 singer]$_2$ helped [her]$_2$ calm down during a thunderstorm.*

La répartition des coréférents dans leurs classes respectives se nomme la résolution des chaînes de coréférences.

Ce travail n'essaiera pas d'unifier les liens des coréférences au sein d'une classe. Il s'agit d'une étape supplémentaire de la résolution dont le succès requiert la résolution de l'ensemble des coréférences, quelques soient leurs natures. L'acquisition et l'exploitation des connaissances nécessaires à la résolution de certaines anaphores, comme les anaphores nominales ou les ellipses, représentent à l'heure actuelle un verrou scientifique important [Boudreau & Kittredge, 2005]. Pour unifier la classe des coréférents de *Sophia Loren* dans notre précédent exemple il faut savoir que *Sophia Loren* est une actrice.

2.1.3 Limitation à l'anaphore pronominale de rang 3

Notre dessein est de mesurer l'apport d'un raisonnement qui accepte les informations quelles que soient leurs qualités dans une tâche d'annotation. Pour que l'apport soit sensible, il nous

[5] L'exemple est emprunté à [Mitkov, 2002].

fallait choisir une tâche qui mette en jeu un grand nombre d'informations dont le calcul automatique soit possible et imparfait. Parmi les tâches d'annotations envisageables nous avons retenu le problème de la résolution des pronoms anaphoriques *it*. Nous excluons intentionnellement les cas où le pronom est cataphorique et où l'antécédent est une entité abstraite. La relation anaphorique du pronom *it*, ainsi limitée, est sans doute l'une des plus simples et des mieux étudiées. Nous allons passer en revue les algorithmes qui découlent de ces études.

2.1.4 Algorithme général pour la résolution automatique des anaphores

La résolution des pronoms *it* anaphoriques consiste à reconnaître les pronoms anaphoriques et à leur adjoindre leur antécédent. On admet habituellement que l'anaphore est correctement résolue si la substitution entre l'anaphore et son antécédent ne produit pas de glissement de sens.

Pour réaliser la résolution automatiquement nous pouvons distinguer trois grandes étapes : la reconnaissance des pronoms impersonnels et anaphoriques, la création de la liste des candidats à l'antécédence et le choix de l'antécédent parmi les candidats. Certains systèmes de résolution confondent certaines étapes mais nous pouvons toujours les distinguer artificiellement. L'algorithme de [Dagan & Itai, 1990] qui s'appuie sur des fréquences de distribution séléctionne l'antécédent en même temps qu'il dresse la liste des candidats. Pour présenter les différentes étapes de la résolution, considérons le contexte suivant :

> *Membrane fusion would be expected to have profound consequences for subsequent development. For example,* $[it]_1$ *is suggested that fusion activates processing of* $[pro - sigmaE]_2$ *to sigma E in the cytoplasm by exposing* $[it]_2$ *to a membrane-bound processing enzyme.*

Ce contexte est composé de deux phrases extraites d'un résumé d'un article scientifique de génomique.

Distinction des pronoms anaphoriques

Durant la première étape tous les pronoms du résumé sont examinés un par un. Il s'agit de distinguer les pronoms anaphoriques des pronoms impersonnels. Un pronom impersonnel est un pronom qui ne renvoie à aucune entité du discours, il est alors sémantiquement vide. Cette distinction annule la recherche inutile d'antécédents aux pronoms impersonnels et les élimine de la liste des candidats à l'antécédence des pronoms anaphoriques. Dans notre contexte il est clair que le pronom 1 est impersonnel alors que le pronom 2 est anaphorique. Le pronom 1 sera ignoré lors des étapes suivantes.

Sélection des candidats

Une fois que les pronoms anaphoriques sont connus, il faut dresser pour chacun d'entre eux la liste des candidats possibles pour l'antécédence. Pour la résolution des anaphores pronominales, il suffit de lister tous les syntagmes nominaux de la phrase où apparaît le pronom ainsi que les syntagmes nominaux des phrases précèdentes. Le nombre de phrases de recherche est à définir en paramètre du système.

Comme le pronom 2 est le seul pronom anaphorique de notre exemple, nous ne devons constituer qu'une seule liste, la liste *C* ci-dessous. Cette liste est uniquement composée des syntagmes nominaux de la phrase du pronom et de la phrase précédente. Pour notre exemple nous prenons la plus petite fenêtre possible.

> *C={Membrane fusion, profound consequences, subsequent development, fusion, processing, pro-sigma E, sigma E, processing of pro-sigma E to sigma E, the cytoplasm}*

Choix de l'antécédent

Lors de la dernière étape le système doit être capable de désigner un candidat dans la liste C. Ce candidat sera l'antécédent proposé par le système pour se substituer au pronom (ici, le pronom 2). Deux stratégies sont possibles et sont souvent mises en oeuvre simultanément.

La première stratégie élimine des candidats de la liste en fixant des contraintes que l'antécédent doit respecter. L'accord de genre et l'accord de nombre entre le pronom et le candidat sont les filtres les plus communnément utilisés et ils éliminent le candidat *profound consequences* de notre liste [6].

La seconde stratégie applique des critères de préférence pour distinguer un candidat particulier dans la liste. Ces préférences peuvent être très simples et facile à calculer comme de choisir le candidat de la liste qui est le plus répété dans le document. Mais ils peuvent aussi être plus complexes comme la recherche du centre d'un contexte anaphorique ([Grosz *et al.*, 1983],[Grosz *et al.*, 1995]).

2.1.5 Mesures d'évaluation

La résolution des anaphores est une tâche complexe qui est encore imparfaitement réalisée automatiquement. De nombreux paramêtres déterminent la réussite de l'algorithme précédent. Cette section détaille les différents paramètres qui doivent être pris en compte pour évaluer les performances d'un système de résolution automatique des anaphores.

Evaluation du filtre des pronoms impersonnels

A notre connaissance, Evans [Evans, 2001] fut le premier auteur a résoudre le problème de la distinction des pronoms impersonnels des pronoms anaphoriques en employant un système de classification automatique. Nous avons suivi la même stratégie dans cette étude, voir 5.1.3). Nous avons donc mesuré l'efficacité de notre système de classification au moyen de l'exactitude.

L'exactitude, *Accuracy (Acc)* en anglais, se définit comme suit :

Acc= $\frac{VP+VN}{VP+VN+FP+FN}$, où

- FP correspond au nombre d'occurrences des pronoms anaphoriques étiquetées comme impersonnelles par le système, nous appellerons ces occurrences les faux positifs (FP),
- FN correspond au nombre d'occurrences des pronoms impersonnels étiquetées comme anaphoriques, nous appellerons ces occurrences les faux négatifs (FN),

[6] L'accord de nombre entre l'anaphore et l'antécédent n'est pas une contrainte sans exception. [Mitkov, 2002] mentionne les deux exemples suivants *The teacher gave [each child]$_1$ a crayon. [They]$_1$ started drawing colouful pictures.* et *Had [the police]$_2$ taken all the statements [they]$_2$ needed from her ?* Dans ces deux exemples les propriétés sémantiques des antécédents sont nécessaires pour démêler le conflit apparent.

– VP et VN correspondent aux nombres d'occurrences des pronoms impersonnels et anaphoriques étiquetées correctement, respectivement, nous appellerons ces occurrences les vrais positifs (VP) et les vrais négatifs (VN).

Evaluation du système de résolution d'anaphores

Critères d'évaluation Il n'existe pas de procédure générale pour évaluer un système de résolution d'anaphores. Plusieurs critères doivent être pris en compte avant de choisir une procédure à défaut d'une autre.

Selon [Mitkov, 2002], le premier critère est d'identifier ce que nous souhaitons évaluer : l'algorithme de résolution seul ou le système dans son ensemble. Lors de l'évaluation de l'algorithme seul, nous évaluons le nombre d'anaphores résolues correctement par l'algorithme lorsqu'il est placé dans des conditions idéales *i.e.* lorsque les informations d'entrée de l'algorithme ne contiennent aucune erreur (elles auront donc auparavant été corrigées). Dans l'évaluation du système de résolution, nous évaluons l'algorithme placé dans les conditions normales d'utilisation. Les informations d'entrée de l'algorithme ont été calculées automatiquement par les outils d'annotation qui composent le système (segmenteur, analyseur syntaxique...) et n'ont subi aucune correction.

Le second critère dont nous devons tenir compte dans l'évaluation est celui de la robustesse du système. Les systèmes, dit robustes, proposent un antécédent pour toutes les anaphores qu'ils ont à résoudre. D'autres systèmes ne proposent aucun antécédent lorsque l'ambiguité est trop élévée. Les mesures classiques de rappel et de précision sont inadaptées pour les systèmes robustes [Aone & Bennett, 1995] :

$$\text{Rappel}=\frac{\textit{Nombre d'anaphores correctement résolues}}{\textit{Nombre d'anaphores identifiées par le système}}$$

$$\text{Précision}=\frac{\textit{Nombre d'anaphores correctement résolues}}{\textit{Nombre de tentatives de résolutions d'anaphores}}$$

Un système robuste propose un antécédent pour l'ensemble des anaphores identifiées par le système. La mesure du rappel d'un système robuste est donc identique à la mesure de précision. De plus la mesure de précision avantage les systèmes non robustes. Ils ne proposent un antécédent que lorsqu'ils sont sûrs du candidat et laissent sans réponse les cas les plus difficiles. Leur précision est normalement meilleure que celle des systèmes robustes.

Pour évaluer un système, il est d'usage de comparer ses performances avec celles obtenues par d'autres systèmes de l'état de l'art sur un corpus de référence. Mais un autre critère doit être pris en compte dans la comparaison des différents systèmes : les informations d'entrée exploitées par les systèmes. Ces informations doivent être similaires. Or, les systèmes sont souvent composés d'outils différents. Les performances globales des systèmes dépendent largement des performances respectives des outils qui les composent. Si les performances des outils diffèrent trop d'un système à l'autre, la comparaison est impossible. De même, sans corriger l'ensemble des informations, certains auteurs choisissent de corriger certaines informations, comme la segmentation des phrases et des mots uniquement, et de laisser inchangées les autres informations. Pour pouvoir être comparé aux systèmes de ces auteurs, le nouveau système devra aussi bénéficier des mêmes corrections.

Procédure d'évaluation L'objectif de notre travail n'est pas de concevoir un nouvel algorithme de résolution des anaphores mais d'évaluer l'impact de l'utilisation d'informations im-

parfaites dans cette résolution. Notre procédure d'évaluation consiste, en conséquence, à comparer les performances de systèmes exploitant des annotations d'entrée imparfaites produites par des outils identiques et dont les algorithmes de résolution utilisent des ensembles choisis d'informations d'entrée.

Tous les systèmes de résolution que nous évaluons dans le chapitre 6 sont des systèmes robustes. Nous avons donc choisi une mesure courante pour ce type de système, le taux de succès :

$$\text{taux de succès} = \frac{\textit{Nombre d'anaphores correctement résolues}}{\textit{Nombre total d'anaphores}}$$

La mesure du taux de succès peut être modifiée pour préciser la difficulté de la résolution de certaines anaphores :

$$\text{taux de succès non-trivial} = \frac{\textit{Nombre d'anaphores correctement résolues - Nombre d'anaphores triviales}}{\textit{Nombre total d'anaphores - Nombre d'anaphores triviales}}$$

$$\text{taux de succès critique} = \frac{\textit{Nombre d'anaphores correctement résolues - Nombre d'anaphores triviales - Nombre d'anaphores non-critiques}}{\textit{Nombre total d'anaphores - Nombre d'anaphores triviales - Nombre d'anaphores non-critiques}}$$

où une anaphore triviale est une anaphore pour laquelle il n'existe qu'un candidat possible et une anaphore non-critique est une anaphore pour laquelle il n'existe qu'un candidat possible après le filtrage du genre et du nombre des candidats possibles. Dans la suite de ce travail nous n'utiliserons pas les mesures des taux de succès non-trivial ou critique bien que nous les ayons calculés. Ils sont identiques au taux de succès lors de nos expériences. Le nombre de candidats possibles pour l'antécédence du pronom neutre *it*, le seul pronom que nous cherchons à résoudre, est toujours supérieur à un (même après application de nos filtres).

En pratique le taux de succès est une mesure qui est difficile à appliquer pour évaluer des systèmes de résolution. Les analyseurs syntaxiques requis pour calculer les syntagmes nominaux produisent rarement une analyse parfaite. Dans la liste des candidats obtenue à partir de cette analyse, certains candidat ne sont identifiés que partiellement ou font défaut. Les performances de l'algorithme de résolution ne peuvent donc pas atteindre 100%. Pour rendre compte de ce phénomène deux mesures reposant sur le taux de succès ont été proposées : le taux de succès strict et le taux de succès partiel.

taux de succès strict $= \frac{\textit{Nombre d'anaphores strictement résolues}}{\textit{Nombre total d'anaphores}}$, où une anaphore est dite strictement résolue seulement si le candidat proposé correspond exactement au syntagme annoté.

taux de succès partiel $= \frac{\textit{Nombre d'anaphores strictement résolues + Nombre d'anaphores partiellement résolues}}{\textit{Nombre total d'anaphores}}$, où une anaphore est dite partiellement résolue si le candidat proposé recouvre partiellement sinon exactement le syntagme annoté.

Considérons par exemple la phrase :

> *[Beta-Galactosidase expression from the spl-lacZ fusion] was silent during vegetative growth and was not DNA damage inducible, but [it] was activated at morphological stage III of...*

Si un système propose *Beta-Galactosidase expression* comme antécédent au lieu du GN entier, l'anaphore est partiellement résolue mais non strictement résolue.

2.2 Les systèmes de résolution automatique d'anaphores

On oppose souvent en TAL les systèmes qui exploitent des connaissances linguistiques et ceux qui reposent sur des indices de surface.

Les premiers systèmes exploitent des connaissances complexes qui, du fait de leur complexité, ne sont pas toujours fiables lorsqu'elles sont calculées automatiquement et peuvent faire défaut ou être incomplètes lorsqu'elles sont produites par l'homme.

Les seconds systèmes s'appuient généralement sur des indices de surface qui sont plus faciles à obtenir mais qui ne permettent de traiter que les cas simples ou les plus courants de la tâche dévolue au système

Sur la tâche particulière de la résolution des anaphores pronominales, en raison de la complexité et du nombre de connaissances nécessaires, l'opposition des systèmes à base de connaissances linguistiques complexes et d'indices de surface a été historiquement très marquée. Les sections suivantes présentent les systèmes de référence pour chaque type d'approche.

2.2.1 Les systèmes exploitant des connaissances linguistiques complexes

Pour chaque étape de la résolution, les premiers systèmes proposés dans la littérature exploitaient des connaissances linguistiques complexes traduisant les contraintes syntaxiques et sémantiques qui régissent l'anaphore. Mais le calcul automatique de ces connaissances était soit impossible, soit trop peu fiable pour être utilisable. Les connaissances linguistiques étaient donc produites manuellement, ce qui présupposait un important travail d'analyse des textes et rendait très difficile l'utilisation automatique de ces systèmes sur des corpus de tailles importantes.

Les stratégies syntaxiques

Parmi les premiers travaux sur la résolution des anaphores le premier algorithme de Hobbs, décrit dans [Hobbs, 1976] et appelé approche naïve de Hobbs (*Hobbs's naïve approach*), reste sans aucun doute l'un des plus célèbres. Cet algorithme résoud les relations anaphoriques pour les pronoms *he, she, it* et *they* ainsi que les pronoms possessifs. Les occurences du pronom *it* impersonnel et les relations anaphoriques clausales pour ce même pronom ne sont pas gérées.

L'algorithme présuppose la connaissance d'une structure syntaxique parfaite. La structure est obtenue par l'application d'une grammaire contextuelle décrite dans l'article. La structure de l'arbre doit correspondre à la structure grammaticale réelle de la phrase, *i.e.* les liens de dépendance des syntagmes sont corrects et tous les syntagmes implicites, comme les anaphores zéro, sont ajoutés dans la structure. L'algorithme parcourt la structure à la recherche d'un syntagme nominal compatible avec le genre et le nombre du pronom.

L'évaluation a été réalisée manuellement sur 300 textes de trois domaines différents : un traité historique, un roman littéraire et des articles journalistiques. Les performances de l'algorithme sont très bonnes avec un taux de succès de 91,7% sur la totalité des anaphores et de 81.8% si on ne considère que les anaphores non-triviales, *i.e.* les anaphores pour lesquels au moins deux candidats pour l'antécédence sont possibles. La qualité des résultats explique en partie le succès de cet algorithme dans la communauté. Il a servi et sert encore de référence dans l'évaluation des systèmes de résolution.

Un autre système s'appuyant sur la syntaxe a influencé les travaux de nombreux chercheurs. Le système *RAP (Resolution of Anaphora Resolution)* [Lappin & Leass, 1994] est dédié à la résolution du pronom anaphorique de 3^{eme}personne. L'algorithme du système repose sur une analyse syntaxique en profondeur calculé par la *Slot Grammar* de [McCord, 1993].

L'algorithme cherche à écarter le maximum de candidats possibles en appliquant trois filtres. Le premier filtre vérifie la cohérence morphologique entre l'anaphore et le candidat. Les deux derniers étudient la structure syntaxique de la phrase du pronom pour déterminer des dépendances syntaxiques entre le pronom et certains candidats pour les rejeter. Le dernier filtre est destiné aux pronoms réflexifs (*himself, myself...*) et réciproques (*each other, one other...*). Pour départager les candidats restants, un poids de la saillance est calculé. Les critères pour calculer la saillance sont ceux de la distance du candidat, de son rôle syntaxique, d'un ordre de préférence Sujet>COD>COI, et de son enchâssement dans les syntagmes nominaux et adverbiaux. Une mise à jour du poids selon la lecture des phrases et la création de classe d'équivalence pour les coréférents permet de tenir compte d'un état d'attention du lecteur dans le choix de l'antécédent.

Le système a été testé sur 345 phrases séléctionnées aléatoirement dans un corpus composé de 48 manuels d'informatique. Les auteurs ont adapté l'algorithme de l'approche naïve de Hobbs à leur analyseur syntaxique pour pouvoir comparer les performances des deux approches. Les résultats globaux reportés sont en faveur du système RAP avec 86% des anaphores résolues contre 82% pour leur implémentation de l'algorithme de Hobbs. De bons résultats qu'il faut toutefois nuancer. Le corpus utilisé pour les tests présente peu d'anaphores interphrastiques, 20% des anaphores sont de ce type. Or, le taux de succès de ce système sur ce type d'anaphore est moindre 74% contre 89% pour l'algorithme de Hobbs. [Preiss, 2001] lors d'une étude de l'impact d'un changement d'analyseur syntaxique sur les performances d'un algorithme, observe une diminution importante des performances de l'algorithme lorsqu'il est appliqué à un corpus d'un autre domaine, le British National Corpus (BNC). Les performances publiées par l'auteur vont de 61% à 64% selon l'analyseur. L'observation est renouvelée par [Qiu *et al.*, 2004] lors d'une implémentation du système RAP avec une performance de 57.9% sur le corpus MUC-6.

Les stratégies sémantiques

Si l'approche syntaxique de Hobbs fournit d'excellents résultats, elle échoue pour certaines résolutions. Dans ce même article [Hobbs, 1976], Hobbs propose un deuxième algorithme pour retrouver les liens sémantiques implicites entre les entités du discours. L'auteur suppose qu'un prétraitement sélectionne les prédicats pertinents des phrases du texte et traduit ces dernières en énoncés du premier ordre. Les prédicats forment les entrées d'un lexique qui encode les connaissances sémantiques du monde. Une série d'inférences révèle les faits implicites du texte et unifie les entités du discours. Commentons un exemple de l'auteur :

The boy walked into [the bank]$_1$. Moments later he was seen on [its]$_1$ roof.

Pour résoudre cette anaphore supposons que le système isole trois objets du discours, x_1 pour la banque, x_2 pour le pronom possessif *its* et x_3 pour le toit. Deux prédicats peuvent décrire les objets précedents, *bank(x_1)* qui se traduit par x_1 est une banque et *roof(x_3,x_2)* par x_3 est le toit de x_2. Pour être en mesure d'unifier les objets x_1 et x_2, le système doit avoir à disposition les deux connaissances suivantes :

– $\forall y(bank(y) \rightarrow building(y))$ qui se lit ainsi toute banque est un bâtiment
– $\forall y \exists z(building(y) \rightarrow roof(z, y))$ qui s'interprète par tout bâtiment a un toit

Le système muni des bonnes règles d'inférence concluera alors que $x_1 = x_2$ apportant ainsi la solution à l'anaphore. Si les antécédents des pronoms anaphoriques ne sont pas trouvés au terme de l'unification, l'algorithme choisit l'antécédent qui a le plus d'occurences dans le discours.

La difficulté de traduire automatiquement une phrase en un énoncé du premier ordre et l'indisponibilité des connaissances sémantiques rendent l'évaluation très délicate. L'algorithme est donc déroulé à la main sur 4 exemples pour s'assurer de son succès, 2 des 4 exemples sont des exemples où l'approche naïve échoue, les 2 autres étant des exemples classiques de la littérature.

Les stratégies reposant sur le discours

Les fondements de la théorie du centre ont été posés durant les années 80. Les travaux de [Grosz *et al.*, 1983] et [Grosz *et al.*, 1995] restent les plus cités. Cette théorie modélise un concept intuitif : le sens d'une proposition repose autour d'un élément central. Grâce à ce concept les auteurs clarifient un ensemble de contraintes que l'orateur est obligé de respecter lorsqu'il emploie une expression coréférentielle.

Pour les auteurs, tout discours a pour but de communiquer des informations, c'est le *Discourse Purpose*. Pour les communiquer, le discours est divisé en segments *Discourse Segment Purpose (DSP)*, où chaque DSP contribue à la réalisation du partage des informations. La cohérence globale du discours dépend des relations des DSP entre eux et de la cohérence de chaque DSP. Cette cohérence locale est conditionnée par les liens entretenus entre les propositions d'un même segment. La théorie du centre permet de modéliser cette cohérence. Plus un interlocuteur a besoin de faire d'inférences pour comprendre l'enchaînement de deux propositions, plus leur cohérence est affaiblie. Si une inférence ne peut aboutir, les propositions sont incohérentes. Nous reproduisons ici deux contextes donnés en exemple par nos auteurs :

– *1A. John went to his favorite music store to buy a piano.*
– *1B. He had frequented the store for many years.*
– *1C. He was excited that he could finally buy a piano.*

– *2A. John went to his favorite music store to buy a piano.*
– *2B. It was a store John had frequented for many years.*
– *2C. He was excited that he could finally buy a piano.*

Le premier contexte est, selon nos auteurs, plus cohérent que le second. A la lecture de la première phrase (1A ou 2A) deux centres sont possibles, John ou le magasin. Alors que le premier contexte focalise l'attention du lecteur sur le centre John tout le long du segment, le deuxième contexte oriente le lecteur sur le magasin puis change brutalement de centre pour John. Ce changement oblige le lecteur à une inférence supplémentaire qui est nécessaire pour résoudre le pronom anaphorique *he* et comprendre le sens du segment, affaiblissant la cohérence du discours.

L'article [Grosz *et al.*, 1995] propose une série de règles pour garantir la cohérence d'un segment du discours. Un segment du discours est constitué par un ensemble de phrases où chaque phrase est composée par une ou plusieurs propositions. Notons U=$U_1, U_2, ..., U_n$ l'ensemble des propositions du segment D. Chaque proposition U_i de U possède un ensemble non vide d'entités

appelées *Forward-looking centers*, notées $C_f(U_i)$, et pour U_i, $i > 1$, une unique entité appelée *Backward-looking center* et notée $C_b(U_i)$. Ces entités sont de nature sémantique et non syntaxique puisqu'il s'agit des dénotations de certains syntagmes nominaux de la proposition U_i [7]. Le *Backward-looking center* est un élément qui appartient à l'ensemble des *Forward-looking centers*. Cette entité est calculée à partir de la proposition précédente U_{i-1} et relie celle-ci à la proposition U_i. Les entités du *Forward-looking centers* sont partiellement ordonées selon l'ordre de saillance (ici calculé grâce au rôle syntaxique). Le premier élement (ou un des premiers élements) de cette proposition est le *Backward-looking center* le plus probable pour la proposition suivante. Considérons le contexte 1 précédent. La première proposition n'a, par définition, aucun *Backward-looking center*. On calcul $C_f(U_1) = \{$*piano, music store, John*$\}$, l'ensemble des objets du discours de *1A*. Les entités de cet ensemble sont ordonnées selon leur rôle syntaxique. Le premier élément de cet ensemble est choisi par l'orateur comme centre de la proposition suivante et on obtient $C_b(U_2) = \{$*John = referent("he")*$\}$.

Deux facteurs peuvent dégrader la cohérence entre 2 propositions. Le premier facteur est l'expression linguistique choisie par l'orateur pour répéter une entité des *Forward-looking centers* ou du *Backward-looking center* dans la proposition suivante car elle influence le nombre d'inférences nécessaires à l'interlocuteur pour interprêter la proposition. Par exemple, le choix d'une relation anaphorique augmente le nombre d'inférences. La règle 1 réglemente l'utilisation des relations anaphoriques : si chaque élément de $C_f(U_i)$ est réalisé par un pronom dans $U_n + 1$ alors le $C_b(U_n + 1)$ doit aussi être réalisé par un pronom. Le second facteur qui peut détériorer la cohérence du segment, est un changement trop brutal du centre d'une proposition à l'autre. Selon les auteurs il existe trois types de changements possibles :

- *continu*, le centre $C_b(U_n)$ est conservé dans $U_n + 1$ et il est aussi le premier élément dans $C_f(U_n + 1)$
- *retenu*, le centre $C_b(U_n)$ est conservé dans $U_n + 1$ mais il n'est plus le premier élément dans $C_f(U_n + 1)$
- *le changement de centre*, le centre $C_b(U_n)$ diffère du centre $C_b(U_n + 1)$

La règle 2 décrit les types de progressions dans les changements du centre. Cette règle ne peut être appliquée sans les critères utilisés pour ordonner les entités des ensembles C_f. Le critère étudié dans [Grosz *et al.*, 1995] est syntaxique, il s'agit d'un ordre de préférence des rôles syntaxiques des syntagmes de la proposition : Sujet>COD>Autres. Nous renvoyons le lecteur interessé par l'étude de ces facteurs aux travaux originaux de l'auteur [Grosz *et al.*, 1995] qui donne et commente un grand nombre d'exemples.

Le cadre dessiné par la théorie du centre permet en dressant la liste des *Forward-looking centers* d'une proposition U_i de réduire le nombre de candidats possibles à l'antécédence et de choisir un bon candidat, normalement le premier élément de cette liste, pour interpréter un pronom anaphorique dans la proposition suivante. Ce cadre fut à l'origine d'un grand nombre d'algorithmes de résolutions d'anaphores dont le plus célèbre est celui de [Brennan *et al.*, 1987].

[7] La détermination des dénotations de chaque syntagme est un mécanisme pouvant nécessiter des facteurs aussi bien sémantiques que pragmatiques. Dans certains discours un syntagme nominal d'une proposition peut être interprété par plusieurs dénotations. [Grosz *et al.*, 1983] analyse un exemple de Kripke où la dénotation du syntagme *her husband* désigne deux entités l'amant et le mari. Les deux entités seront ajoutées à l'ensemble des *Forward-looking centers* de la proposition.

Des stratégies difficiles à appliquer automatiquement

Les évaluations manuelles de ses stratégies servent, aujourd'hui, de références pour estimer la qualité des outils produits pour la résolution des anaphores et de mesurer leur progrès. En effet, la précision et la complétude des connaissances à disposition des systèmes de TAL restent encore insuffisantes pour appliquer ces stratégies automatiquement avec succès. La production d'une analyse syntaxique complète d'une phrase, le choix des prédicats pertinents et l'inférence nécessaire à la résolution ou encore le calcul de la cohérence d'un contexte local requis pour déterminer l'élément saillant de ce contexte sont toujours des problèmes difficiles que nous ne savons que partiellement résoudre. La section suivante présente les différentes solutions proposées pour pallier l'absence de ces connaissances.

2.2.2 Les systèmes reposant sur des indices de surface

Durant les années 1990, devant le besoin de systèmes de résolutions robustes et peu coûteux à mettre en place, un nombre important de systèmes à base d'indices de surface ont été proposés [Mitkov *et al.*, 2001]. Ces systèmes abandonnent les connaissances linguistiques complexes des premiers systèmes. Ils approchent les connaissances nécessaires par des indices plus simples et que l'on suppose plus fiables.

Les nombreux travaux existants proposent des systèmes différents pour sélectionner le meilleur candidat en exploitant bien souvent des ensembles plus ou moins identiques d'indices de surface. Nous ne nous attarderons pas sur les systèmes en question, hormis le système MARS en fin de section car nous l'utiliserons comme système de comparaison dans nos expériences. Nous préférons dresser une liste, certainement incomplète, des connaissances linguistiques que les auteurs ont jugé pertinents à utiliser et les heuristiques qu'ils ont imaginées pour les approcher. Cette liste nous servira de base pour établir le choix des informations que nous modéliserons pour notre système de résolution dans la section 6.

Informations Distributionnelles

Les patrons de collocations Devant l'indisponibilité des informations sémantiques (voir la section 2.2.2), [Dagan & Itai, 1990] proposent une méthode distributionnelle pour approcher ces connaissances sémantiques. Le principe linguistique sous-jacent est simple. Il repose sur l'hypothèse que l'antécédent et le pronom anaphorique partagent les mêmes contextes. Pour lever l'ambiguïté entre deux candidats, [Dagan & Itai, 1990] comparent les contextes de chacun des candidats au contexte du pronom. Le candidat qui a le plus grand nombre de contextes similaires à celui du pronom est choisi comme antécédent.

Une analyse syntaxique en dépendance du corpus est nécessaire pour calculer les contextes. [Dagan & Itai, 1990] listent toutes les relations sujet-verbe et objet-verbe présentes dans le corpus. Lorsqu'un pronom est en relation de sujet ou d'objet d'un certain verbe, l'antécédent est le candidat qui apparaît le plus de fois dans la même relation que le pronom avec le verbe. [Dagan & Itai, 1990] donnent l'exemple suivant :

> *They knew full well that the companies held tax [money]$_1$ aside for collection later on the basis that the [government]$_2$ said [it]$_2$ was going to collect [it]$_2$.*

money est l'antécédent du deuxième pronom car *money* est 149 fois complément d'objet du verbe *collect* dans le corpus alors que *government* n'est jamais complément d'objet de ce verbe.

L'analyse syntaxique en dépendance d'un document est difficile à obtenir automatiquement. Son calcul est très bruité surtout si la langue du corpus est une langue de spécialité [Aubin *et al.*, 2005]. [Mitkov, 1994] proposent une première heuristique qui substitue à l'analyse syntaxique en dépendance un simple calcul des SN. Cette heuristique sera aussi utilisée dans [Boudreau & Kittredge, 2005]. Ce calcul s'effectue au moyen de règles écrites manuellement à partir de la grammaire de la langue. Les patrons distributionnels sont de fait simplifiés et de la forme : <*SN/Pronom Verbe*> et <*Verbe SN/Pronom*>. L'étude de [Kehler *et al.*, 2004] relève une autre difficulté classique à la distribution. Elle est calculée sur le corpus. Or toutes les combinaisons acceptables ne sont pas nécessairement observées dans le corpus. Pour réduire le silence distributionnel notre auteur propose, en s'appuyant sur les travaux de [Keller & Lapata, 2003], une seconde heuristique où les distributions sont calculées sur le Web.

Répétition des candidats La répétition des candidats est une information distributionnelle incontournable pour les systèmes d'apprentissage pauvres en connaissances, et qui est citée par exemple dans [Ge *et al.*, 1998]. Cette information privilégie le candidat qui est le plus souvent répété dans le document pour le choix de l'antécédent :

> *Jenny has spotted a nice [cup]. She bought the coveted [cup]. However, once back home, she put [the cup]*$_1$ *on a plate and broke [it]*$_1$*...*

Ce choix permet un calcul rapide et simple des éléments saillants du document.

Le plus simple pour repérer ces répétitions est de comparer les chaînes des mots du document. Le calcul de la répétition des candidats peut, toutefois, être plus sophistiqué [Mitkov, 2002]. Nous pouvons rechercher puis comptabiliser les synonymes des candidats, ou encore, les SN où le candidat apparaît en tête (*the bottle* est une répétition de *bottle of toner*). Des SN ayant une tête identique seront ajoutés au total même s'ils ne sont pas coréférentiels entre eux comme *the first channel* et *the second channel*.

Distance entre le candidat et le pronom La distance entre un candidat et un pronom est, avec la répétition des candidats, une information très souvent employée. Elle permet d'apprécier simplement la récence du candidat à l'esprit du lecteur : plus le candidat est loin du pronom moins ce dernier a de chance d'être l'antécédent. La distance est couramment utilisée comme heuristique dans le calcul du centre. Plusieurs métriques peuvent être choisies. La plus simple consiste à mesurer le nombre de phrases de distance entre le pronom et le candidat [Soon *et al.*, 2001] ; 0 s'ils sont dans la même phrase, 1 si le candidat est dans la phrase précédente... Cette mesure peut être affinée en comptant la distance en terme de propositions plutôt que de phrases. Une mesure plus sophistiquée reportée par [Ge *et al.*, 1998] compte le nombre de GN rencontrés par l'algorithme de Hobbs avant qu'un candidat acceptable soit trouvé.

Informations Morphologiques

Genre, Nombre Les informations morphologiques du genre et du nombre[8] servent habituellement de contraintes pour la sélection d'un candidat. Si le genre ou le nombre du candidat

[8] On classe parfois ces informations comme des informations sémantiques.

sont différents de celui du pronom, le candidat est rejeté. Le critère du nombre n'est cependant pas parfait. Certains candidats singuliers peuvent être repris par un pronom pluriel comme le montre l'exemple de [Ge *et al.*, 1998] :

I think if I tell [Viacom]$_1$ I need more time, [they]$_1$ will take 'Cosby' across the street

Aussi simples que soient ces informations, leurs calculs posent, encore aujourd'hui, problème. Le calcul du nombre d'un SN composé de plusieurs mots demande d'avoir isolé la tête du syntagme pour en déterminer le nombre [Cardie & Wagstaff, 1999]. Lorsque le syntagme est un unique mot, une analyse morphologique de celui-ci suffit. La reconnaissance du genre d'un SN est plus compliquée. [Soon *et al.*, 2001] s'en remettent aux titres tels que *Mr.*, *Mrs.* pour s'assurer du genre d'un syntagme. Si aucun titre ne précède le syntagme, on le recherche dans un dictionnaire des noms propres. Pour les syntagmes qui ne sont pas des personnes, les auteurs induisent le genre des syntagmes grâce à leur classe sémantique trouvée en interrogeant *WordNet*. [9].

Pronom Certains systèmes distinguent les pronoms lorsqu'ils sont en position de candidat pour l'antécédence d'un autre pronom. Le centre du discours peut être marqué par des reprises anaphoriques successives comme dans le contexte suivant :

> *CSF has at least three activities : (i) at low concentrations, it stimulates expression of genes activated by the transcription factor ComA ; at higher concentrations, it (ii) inhibits expression of those same genes and (iii) stimulates sporulation.*

Le système doit alors être capable de restituer les informations sémantiques du premier pronom pour que le second puisse en hériter.

Nom Propre Tout comme les pronoms, les noms propres peuvent être singularisés par les systèmes. Les systèmes que nous avons cités réduisent la notion aux prénoms et aux noms de famille uniquement, mais cette notion peut être étendue selon le domaine du corpus. Les noms de gènes peuvent ainsi jouer le rôle de noms propres dans un texte de génomique. Les méthodes pour identifier les noms propres vont évidemment varier selon les entités nommées. Pour retrouver les noms propres associés à des personnes nous pouvons à nouveau rechercher les titres dans le contexte ou une entrée dans un dictionnaire. [Cardie & Wagstaff, 1999] proposent une heuristique encore plus simple puisqu'elle s'appuie uniquement sur les majuscules des lettres initiales des mots. La casse est aussi un bon indicateur pour retrouver les noms de gènes ou de protéine (voir par exemple [Proux *et al.*, 1998]).

Informations Syntaxiques

La tête du candidat La tête du syntagme nominal n'apparaît pas comme information dans la représentation des systèmes de résolution d'anaphore pronominale mais son calcul est nécessaire pour calculer les valeurs d'autres informations comme le nombre du syntagme. Le calcul

[9] Le lecteur intéressé pourra trouver des informations complètes concernant cette ressource sur le site web : `http://wordnet.princeton.edu`.

est souvent confié à un analyseur syntaxique en profondeur et, lorsque ce dernier est indisponible, une heuristique comme celle de [Cardie & Wagstaff, 1999] (retenir le dernier du mot du SN comme tête du candidat) peut être utilisée pour l'approcher.

Candidat défini, démonstratif ou possessif La nature de l'article qui précède le syntagme est aussi une des informations utilisées dans les systèmes reposant sur une appréciation du centre. Un candidat défini ou précédé d'un article démonstratif est mis en valeur par l'orateur, contrairement à un candidat indéfini. Une simple analyse du contexte du candidat permet d'identifier la présence d'un article et de classer le candidat en conséquence.

Candidat appositif La position particulière d'un syntagme mis en apposition peut être un indicateur d'une moindre saillance de ce dernier. Utilisée dans la résolution des anaphores nominales, cette position peut aussi être utilisée comme critère de préférence dans la résolution des anaphores pronominales. Dans l'exemple *Bill Gates, the chairman of Microsoft Corp.*,... les deux syntagmes nominaux sont coréférents. Mais la position particulière du syntagme en apposition est un indicateur du moindre intérêt accordé à ce syntagme par l'orateur. Deux critères sont employés par [Soon *et al.*, 2001] pour deviner si un syntagme est dans une apposition : l'encadrement du syntagme par un signe de ponctuation double et l'absence de verbe entre les ponctuations.

SN prépositionnel Une autre position particulière du syntagme dans la proposition a été utilisée par [Mitkov, 2002] pour pénaliser ce syntagme. Les syntagmes prépositionnels sont, tout comme les syntagmes en apposition, moins saillants dans le discours comme le montre l'exemple

Insert the cassette into the VCR making sure it is suitable for the length of recording

où le candidat *the VCR* est rejeté en raison de sa position. A notre connaissance aucune heuristique n'a été proposée pour cette information, le calcul s'effectue donc sur l'arbre syntaxique complet de la phrase.

Le candidat sujet Le rôle syntaxique d'un syntagme est un bon indicateur selon la théorie du centre. Le thème le plus important de la phrase est souvent introduit par le sujet. Le syntagme qui a le rôle de sujet est donc fréquemment préféré aux autres syntagmes, une préférence qui justifie l'ordre de la section 2.2.1 :*Les stratégies reposant sur le discours*. Toutefois le rôle grammatical des syntagmes est difficile à calculer automatiquement. En l'absence d'une analyse syntaxique, [Mitkov, 2002] approche le sujet en le confondant avec le premier syntagme de la proposition.

Le parallélisme syntaxique La préférence pour le sujet est un cas particulier du parallélisme syntaxique [Ng & Cardie, 2002]. Dans le cadre du parallélisme syntaxique on préfère le syntagme qui à le même rôle syntaxique que le pronom dans certaines conditions [Carbonell & Brown, 1988]. Par exemple dans le cas où le pronom apparaît dans une proposition coordonnée, on préfèrera le syntagme de la proposition précédente qui a le même rôle grammatical que le pronom :

The programmer successfully combined [Prolog]$_1$ with C, but he had combined [it]$_1$ with Pascal last time.

Ici on préfère le COD au sujet pour l'antécédence du pronom.

Cataphores Peu de systèmes affrontent le problème de la résolution des cataphores, ce phénomène linguistique étant plus rare que l'anaphore. Mentionnons toutefois la tentative de [Rich & LuperFoy, 1988] pour les résoudre au moyen de règles reconnaissant les constructions syntaxiques introduisant des cataphores comme

When he is happy, John sings.

Informations Sémantiques

Candidat animé Parmi les travaux que nous avons cités, les auteurs répartissent les syntagmes en deux classes distinctes : les objets animés comme les hommes ou les animaux et les objets inanimés (les tables, les chaises, *etc*). La distinction a, jusqu'à présent, été réalisée grâce à l'interrogation des définitions fournies par une unique ressource : *WordNet*.

Verbes spécifiques Certains verbes focalisent l'attention du lecteur sur le syntagme qui les suit. Ce syntagme est alors l'antécédent. [Mitkov, 2002] a jeté les bases de cette classe sémantique à partir d'une étude d'un corpus de textes techniques. Voici une partie des verbes qu'il donne en exemple :

analyse, assess, check, consider, cover, define, describe...

Terme du domaine La définition de ce que peut être un terme est toujours en discussion au sein de la communauté terminologique [Bourigault & Jacquemin, 2000]. Pour notre utilisation nous pouvons nous satisfaire de la définition de *la théorie générale de la terminologie* :

un terme est un représentant linguistique d'un concept dans un domaine de connaissance.

En tant que représentant d'un concept, les termes sont des éléments importants du discours et en sont souvent des éléments saillants [Mitkov, 2002]. L'acquisition des termes peut se faire grâce à l'exploitation des ressources existantes ou par un outil de construction (semi)automatique sur un corpus de référence. Le lecteur intéressé par les techniques mises en œuvre par les outils de terminologie pourra se reporter à [Bourigault & Jacquemin, 2000].

Cohérence sémantique Dès les premiers travaux sur la résolution des anaphores, les auteurs ont pris conscience de l'insuffisance des informations syntaxiques et du nécessaire recours aux informations sémantiques [Carbonell & Brown, 1988], [Hobbs, 1976]. Examinons les deux contextes de [Carbonell & Brown, 1988] :

- *John take [the cake]$_1$ from the table and ate [it]$_1$*
- *John take the cake from [the table]$_1$ and washed [it]$_1$*

Les deux phrases sont identiques syntaxiquement, elles diffèrent uniquement par les deux derniers verbes et la référence anaphorique. Dans la première phrase le pronom ne peut pas référer au syntagme *table* car *John* est le sujet du verbe *ate* et qu'un homme ne mange, *a priori*, pas

les tables et préfère les gâteaux. Grâce à nos connaissances sémantiques nous formulons un ensemble de contraintes qui nous permettent de lever l'ambiguïté entre les deux candidats concurrents. Après avoir supposé que le pronom réfère au syntagme *the cake*, il hérite des propriétés sémantiques de l'objet *cake*, en particulier qu'il s'agit de nourriture. La classe sémantique d'un syntagme peut aussi être un indice de saillance pour des corpus spécialisés (les gènes peuvent être priviligiés au dépend des noms de personne dans un corpus de génomique). Malheureusement les ressources ontologiques que réclament les systèmes voulant inférer ces contraintes sémantiques sont souvent indisponibles, partielles ou inadaptées [Gandon, 2002], ce qui nuit à leur efficacité. Diverses heuristiques ont été proposées pour supplér au manque de ressources. L'une d'entre elles est l'utilisation des patrons de collocations (voir la section 2.2.2). A défaut d'une ontologie adaptée les relations et les propriétés des *synsets* de *WordNet*[10] sont souvent interrogées pour retrouver la classe et les propriétés sémantiques d'un syntagme [Soon *et al.*, 2001]. Des ressources plus spécifiques ont été récemment développées et peuvent maintenant bénéficier à la résolution des anaphores, c'est le cas par exemple des schémas prédicat-argument (Semantic Role Labeling) de PropBank [Ponzetto & Strube, 2006].

Informations pragmatiques

La compréhension d'un texte dépend fortement du contexte dans lequel nous effectuons la lecture. Les informations que nous réunissons dans cette section dépeignent les aspects documentaires ou logiques du contexte de lecture d'un document ou d'un corpus.

Titre et tête de section Les informations relatives aux documents comme le titre du document ou les têtes de sections sont, du fait de l'augmentation des documents structurés, de plus en plus facilement accessibles aux systèmes. Ces informations qui précisent le sujet de la section à venir, mettent en relief des syntagmes nominaux importants du discours et concourrent à la reconnaissance des éléments saillants d'un contexte [Mitkov, 2002].

L'accessibilité logique La Théorie des Représentations Discursives (DRT) dont les fondements ont été posés par Kamp et Reyle [Kamp & Reyle, 1993], est une représentation logique du sens du discours. Selon cette théorie le sens d'un discours peut être construit par la mise à jour successive d'un contexte représenté par les Structures de Représentation Discursive (DRS). Un algorithme travaillant à partir d'une analyse syntaxique en constituants déclenche un ensemble de règles de construction pour mettre à jour une liste des objets du discours et des relations qu'ils entretiennent. Lorsqu'une anaphore est rencontrée, le contexte courant de la DRS formalise des contraintes d'accessibilité à certains objets du discours qu'un système de résolution utilise pour éliminer certains candidats [Amsili & Bras, 1998]. Les discours

1. *Pedro possède un âne. Il le bat.*
2. *Pedro ne possède pas d'âne. Il le bat.*

sont représentés par les deux DRS de la figure 2.1. Dans la première DRS les objets X et Y qui représentent respectivement Pedro et l'âne sont accesssibles lors de la lecture de la seconde

[10] Dans l'architecture de *WordNet* les mots sont regroupés par ensembles de synonymes appelés *synset*. Les *synset* entretiennent différentes relations sémantiques et lexicale comme l'hyperonymie, l'antonymie... Pour plus de détail voir `http://wordnet.princeton.edu/man/wngloss.7WN`

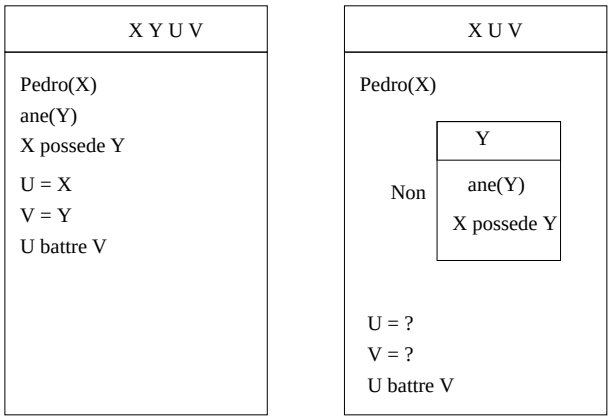

FIG. 2.1 – Deux exemples de la relation d'accessibilté de la DRT

phrase du discours. Le choix du bon antécédent pour les pronoms reste à la charge du système de résolution. Ce dernier est libre d'utiliser des informations syntaxiques et sémantiques pour parvenir à la résolution attendue qui s'exprime dans ce formalisme par les égalités $U = X$ et $V = Y$. Dans la deuxième DRS, l'introduction de la négation rend l'objet Y inaccessible. Il est donc impossible au système d'unifier les objets Y et V et il laisse les égalités incomplètes.

Connaissances physiques et psychologiques du monde Les connaissances que nous avons du monde physique dans lequel nous évoluons (ou que nous imaginons) et des principes psychologiques qui régissent les relations humaines aussi simples soient-elles sont difficiles à représenter formellement. Or [Carbonell & Brown, 1988] notent qu'elles sont nécessaires pour la résolution de certaines anaphores comme celles que les auteurs donnent en exemple :

John gave Tom$_1$ an apple. He$_1$ ate the apple.

Cette anaphore ne peut être résolue que si nous savons représenter le sens du verbe donner et d'inférer que John n'ayant plus la pomme, seul Tom peut la manger. Des travaux plus récents comme ceux de [Nugues *et al.*, 2004] et [Nouioua, 2007] constituent des avancés pleines de promesses pour la modélisation et l'utilisation de ces connaissances. Ces travaux en se limitant au monde fermé des accidents de la route que décrivent les constats d'accident des assurances, rendent possibles la "compréhension" automatique de ces textes par des moteurs d'inférences logiques dédiés ou des simulations graphiques de la scène, que le système pourra certainement à terme interroger pour résoudre les relations anaphoriques non résolues par les autres méthodes.

2.2.3 Le système MARS

Parmi les systèmes pauvres en connaissances présents dans l'état de l'art, nous avons choisi le système MARS dont l'implémentation et l'évaluation sont détaillées dans [Mitkov, 2002]. Les raisons qui justifient notre choix sont tout d'abord le succès de ce système, qui est maintenant une référence pour la communauté du TAL. Ensuite, la stratégie de recherche et de décision de l'antécédent employée par l'algorithme est facilement implémentable. Enfin, l'architecture modulaire du système offre la possibilité d'être enrichie par de nouvelles connaissances.

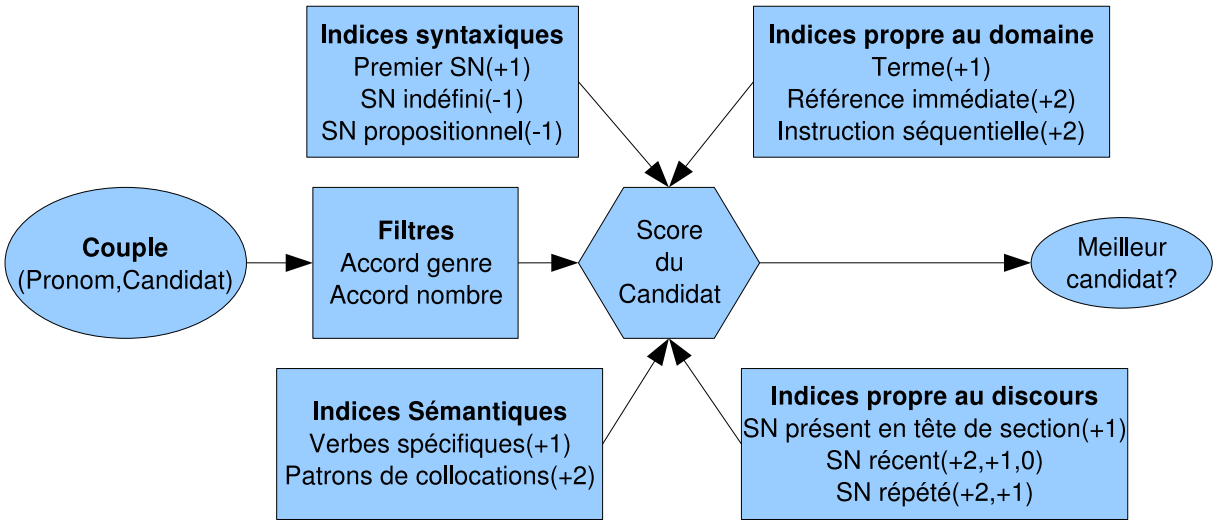

FIG. 2.2 – Architecture du système MARS

Stratégie et architecture du système

Le système MARS, dont la figure 2.2 schématise l'architecture, repose sur une hypothèse simple : le centre est souvent confondu avec l'élément saillant qui précède le pronom anaphorique à résoudre. En choisissant le candidat le plus saillant dans le contexte du pronom qui est bien plus facile à calculer automatiquement que le centre, le système a de fortes chances de résoudre l'anaphore correctement. Comme nous l'avons vu dans la section précédente 2.2.2, de nombreux indices permettent d'influer sur l'importance d'un candidat pour le lecteur. La méthode la plus simple pour associer une saillance à un candidat consiste à associer un poids à ces indices, et de retenir comme antécédent le candidat qui a le plus haut score. C'est la méthode retenue par [Mitkov, 2002] dont nous décrivons ici la version initiale du système MARS [Mitkov, 1998] et qui a été améliorée par la suite dans [Mitkov *et al.*, 2002].

Le système entend résoudre les pronoms personnels, possessifs, démonstratifs et réflexifs. Après avoir ségmenté les textes, le système les analyse avec un *POS tagging* et au moyen des règles simples de grammaire, il dresse la liste des SN dans les deux phrases précèdant un pronom donné. Pour chaque SN associé au pronom, un ensemble d'indices est appliqué : les contraintes et les préférences. Les contraintes éliminent les GN qui ne peuvent être des antécédents[11], les préférences identifient l'antécédent parmi les GN restants. Chaque indice possède un score positif ou négatif et lorsqu'il s'applique au candidat on ajoute son score au score global du candidat. L'antécédent est le candidat qui obtient le plus grand score. En cas d'égalité entre des candidats, le système compare les scores des candidats pour un sous-ensemble ordonné d'indices[12]. Si un candidat obtient un meilleur score que son concurrent pour un indice du sous-ensemble, il est choisi comme antécédent. Si l'égalité subsiste (*i.e.* les candidats ont les

[11] Dans la première version, l'auteur ne parle pas de la méthode employée pour éliminer les pronoms impersonnels, nous supposons qu'ils ont été enlevés manuellement. Dans la deuxième version, ces pronoms seront distingués en utilisant un classifieur identique à celui de [Evans, 2001].

[12] Les indices du sous-ensemble sont les indices *référence immédiate, patron de collocation identique, verbe spécifique*.

mêmes scores pour chacun des indices du sous-ensemble), le système retient le candidat le plus récent.

Deux filtres sont appliqués à l'ensemble des candidats séléctionnés par le système. Le premier est le filtre du nombre. Le filtre reconnaît les groupes nominaux dont le nombre diffère de celui de l'anaphore, par exemple, *government, team, parliament* peuvent être repris anaphoriquement par le pronom *they* ou encore *data* qui peut être l'antécédent du pronom *it*. Le système repère ces noms grâce à un dictionnaire et ne les exclut pas de la liste des candidats possibles. Le second filtre est celui de l'accord du genre entre le pronom et le candidat.

Le système calcule le score de chaque candidat retenu. Les critères de sélection des indices pour le système MARS sont les critères de fiabilité et de simplicité du calcul des indices. A chaque indice est attribué un score qui a été estimé selon les observations des auteurs sur le corpus de développement et que nous reportons brièvement ici :

- Verbes spécifiques (+1)
- Le premier syntagme de la phrase (+1)
- Candidat propositionnel (-1)
- Répétition du candidat dans le titre ou une tête de section (+1)
- Candidat Récent (+2 si le candidat est dans la proposition précédant le pronom, +1 dans la phrase précédente, 0 pour la dernière phrase)
- Candidat indéfini (-1) : les SN qui ne sont pas précédés d'un article défini, possessif ou démonstratif sont considérés comme définis
- Terme du domaine (+1) : les candidats qui contiennent au moins un mot du corpus ayant les meilleurs scores tf-idf[13] sont considérés comme termes du domaine
- Répétition du candidat (+2 pour les candidats qui sont répétés plus de 2 fois dans le paragraphe, +1 pour les candidats répétés seulement une fois) : un candidat est répété si sa chaîne de caractères est retrouvée dans la partie du document précédant le candidat ou si sa tête est répétée dans un syntagme précédent
- Patron de collocations identiques (+2) : en l'absence d'analyse syntaxique en dépendance les patrons sont de la forme <*SN/Pronom Verbe*> et <*Verbe SN/Pronom*> où SN/Pronom désigne le premier syntagme nominal ou le premier pronom précédant (resp. suivant) le verbe *Verbe*. Le système tient compte aussi de la forme négative des verbes (ex. unexpect) avec des patrons de la forme <*SN/Pronom unVerbe*> et <*unVerbe SN/Pronom*>
- Référence immédiate (+2) et Instruction séquentielle (+2) : ces deux indices sont spécifiques au domaine du corpus à partir duquel le système a été conçu et sont malaisément exportables à d'autres domaines. Certaines constructions syntaxiques rencontrées dans les textes techniques mettent en relief un groupe nominal du schéma qui est très probablement l'antécédent d'un pronom anaphorique *it*. Les constructions pour la référence immédiate sont de la forme *(You) Verbe SN ... conjonction (you) Verbe it* où *conjonction* appartient à l'ensemble *and/or/before/after/until...* et les constructions pour les instruc-

[13] Le score tf-idf d'un mot repose sur un poids du mot dans le corpus que l'on calcule à partir d'une loi de répartition des mots dans ce corpus. La loi employée pour le score tf-idf est la loi de Zipf qui suppose que les mots les plus importants d'un document sont les mots représentatifs du corpus, *i.e.* les mots qui n'apparaissent ni trop souvant dans le corpus, comme les mots outils (articles, mot de liaison...), ni trop peu souvent, comme les mots spécifiques à un document précis. Pour une présentation complète du calcul des scores tf-idf, voir [Manning & Shutze, 1999].

tions séquentielles sont de la formes *To Verbe SN, Verbe SN. To Verbe it, Verbe NP.*[14]

Evaluation du système

Le système a été évalué de deux manières différentes. La première évaluation publiée dans [Mitkov, 1998] a permis de mesurer les performances de l'algorithme uniquement. Dans une seconde publication [Mitkov *et al.*, 2002] qui fait état d'une implémentation complète du système, le système entier (*i.e.* analyseur syntaxique+classifieur pour les pronoms impersonnels+ algorithme $1^{ère}$ version) a été évalué.

Pour l'évaluation de l'algorithme l'auteur a travaillé sur un petit corpus composé de textes extraits de manuels techniques qui contient 223 pronoms dont 167 ont été rejetés étant déictiques ou impersonnels. L'évaluation a été réalisée manuellement pour pouvoir corriger les erreurs dans les entrées du système. Les performances annoncées avec un taux de succès de 89.7% sont comparables avec celles des algorithmes s'appuyant sur des connaissances morpho-syntaxiques comme l'algorithme de Hobbs.

L'évaluation du système entier a été effectuée sur un corpus appartenant au même domaine mais contenant plus de pronoms : 2263 pronoms dont 408 sont impersonnels. Sans correction apportée aux entrées, les performances du système sont plus faibles que celles de l'algorithme. En raison des erreurs de l'analyse syntaxique en constituants sur laquelle la liste des candidats est calculée, certains GN candidats ne sont identifiés que partiellement ou font défaut et les performances du système ne peuvent atteindre 100%. Le taux de succès est dit strict lorsque l'antécédent exact a été annoté par le système et partiel lorsque seule une partie contenant la tête de l'antécédent à été annotée. Les résultats sont donnés en taux de succès partiel. Les résultats varient selon le texte considéré : les performances vont de 33% pour les moins bonnes à 64% pour les meilleures, avec une moyenne de 45.81%. La deuxième version du système où des indices ont été ajoutés obtient de meilleures performances avec une moyenne de 59.35%.

Les limites du système

Les travaux de Mitkov ont montré la possibilité de créér un système robuste et utilisable de résolution des anaphores sur des textes réels. Mais, tout comme l'auteur, nous pouvons nous interroger sur la pertinence des scores attribués aux indices. Ces derniers ne sont en effet fondés que par l'observation humaine sur le corpus de développement. L'auteur a tenté d'optimiser les scores des indices de la seconde version du système MARS en recourrant aux méthodes d'apprentissage (un perceptron et un algorithme génétique). Les performances du système optimisé avec un perceptron sont décevantes car légèrement plus faibles que celles du système non-optimisé. En revanche celles du système optimisé avec l'algorithme génétique marquent une amélioration de 2 points sur la moyenne du système (soit une performance de 61.55%).

Pour défendre la stratégie de son système l'auteur avance qu'un indice seul peut indiquer un mauvais candidat comme antécédent, mais que la combinaison des indices permet de corriger cette erreur et de trouver le bon antécédent. Le problème rencontré par le système est celui

[14] Voici deux exemples de cette construction données par l'auteur :
– *To print the paper, you can stand up the printer up or lay it flat*
– *To turn on the video recorder, press the red button. To programme it, press the programme key*

d'une prise de décision en s'appuyant sur des informations imparfaites. Or, la représentation choisie par l'auteur est la plus simple possible et, compte tenu des difficultés éprouvées par l'auteur pour justifier les poids associés aux indices, elle semble inadaptée pour gérer cette imperfection. Différentes réflexions ont été menées pour intégrer l'imperfection d'informations syntaxiques et sémantiques dans la décision, notamment [Mitkov, 1997] ou encore [Ge *et al.*, 1998].

2.3 Conclusion

Le phénomène linguistique de l'anaphore est complexe. La relation anaphorique varie selon différents critères :

- la langue naturelle étudiée,
- le sous-langage étudié,
- l'existence de l'objet dénoté par l'antécédent,
- la nature grammaticale de l'antécédent et de l'anaphore,
- la position de l'antécédent relativement à celle de l'anaphore.

En vue de répondre à notre objectif nous avons réduit le problème général de la résolution anaphorique à un type précis de relation : la résolution du pronom anaphorique anglais *it*.

Cette relation reste probablement l'une des relations anaphoriques les plus faciles à résoudre. Elle a déjà été longuement étudiée et présente les caractéristiques que nous recherchons. Les stratégies proposées par l'état de l'art peuvent être réparties en deux catégories : les stratégies fondées sur des connaissances linguistiques complexes et les stratégies établies sur des indices de surface. Nous avons dressé la liste des informations employées par les systèmes de l'état de l'art les plus célèbres. Cette étude montre que si le calcul des indices de surface est facile et relativement sûr, l'exploitation de ces indices ne suffit pas à résoudre toutes les anaphores pronominales et les cas ambigus. Le recours aux connaissances linguistiques complexes est nécessaire même si elles sont difficiles à calculer automatiquement.

Notre travail consiste donc à concevoir et à implémenter un système de résolution d'anaphores pronominales capable de bénéficier de toutes les sources d'informations disponibles quelle que soient leurs qualités. L'évaluation de notre système mettra en évidence l'apport des annotations imparfaites dans la résolution.

Le système MARS que nous avons présenté dans cette section est proche du système que nous recherchons. Nous nous en inspirons pour concevoir l'architecture de notre propre système. Mais selon nous, il emploie une représentation rudimentaire de l'imperfection des informations qu'il exploite, et surtout, il se limite à des indices de surface pour déterminer l'élément central d'un contexte anaphorique. Deux faiblesses que nous allons chercher à corriger.

Chapitre 3

L'inférence à partir d'informations imparfaites

Sommaire

Au terme du premier chapitre nous avons mis en relief la difficulté pour un système de travailler avec des informations de parfaite qualité. Pour annoter un document le système doit raisonner sur des objets dont la description peut être lacunaire et erronée. Il doit, de plus, être capable d'adapter son raisonnement en fonction du domaine et du genre de corpus. Nous avons présenté une modification de la représentation objet classique des documents. Cette dernière intègre une mesure des degrés de confiance que le système peut avoir sur un ensemble d'annotations concurrentes. Ce chapitre apporte une définition plus précise de cette mesure de confiance et propose un mécanisme de raisonnement apte à tirer profit de cette mesure.

3.1 Des informations imprécises

Dans les exemples des sections précédentes les informations mises à disposition de notre système étaient toujours précises. Le système était toujours capable de déterminer univoquement la valeur d'un attribut d'une annotation. Or l'attribution d'une valeur univoque n'est pas toujours possible.

Selon [Bouchon-Meunier, 1995] deux raisons distinctes peuvent expliquer l'emploi d'informations imprécises. La première est due à notre incapacité à préciser des concepts qui attendent une valeur exacte. Pour certains objets, une description partielle est suffisante. Prenons l'exemple suivant :

La manifestation à Paris était importante,

Si l'adjectif *important* laisse entendre une manifestation réunissant plus de 500 000 personnes, le compte exact à peu d'importance pour les acteurs politiques ou médiatiques et il n'est jamais publié. Pour d'autres objets une description exacte est impossible à produire car le calcul, l'observation ou la mesure est irréalisable comme pour la douleur d'un patient ou la distance entre deux corps celestes.

La seconde raison est inhérente à la nature même de certains concepts. L'appartenance d'un objet à un concept dont les contours sont mal définis est nécessairement imprécise. Considérons les concepts de *chaud* et *tiède*. Existe-t-il un seuil de température pour différencier l'eau chaude de l'eau tiède ? Un bain est dit chaud ou tiède selon les goûts d'une personne, de la saison (un bain à 40° sera chaud en été alors qu'il sera tiède en hiver), de l'âge de la personne (un bain chaud pour un nourrisson est tiède pour un adulte), *etc*... La détermination de ces propriétés pour les objets du monde dépend de leur définition mais aussi du contexte environnant l'objet.

La théorie des sous-ensembles flous a été définie par [Zadeh, 1965] pour représenter ces concepts imprécis. Dans cette théorie un objet peut appartenir entièrement ou seulement partiellement à un ou plusieurs concepts complémentaires. L'appartenance d'un objet à un concept est définie par la fonction caractéristique de ce concept. Formellement, soit X l'ensemble des objets considérés et A un concept,

$x \in A$ ssi $\Delta_A(x) = 1$ où Δ_A est la fonction caractéristique de A : $\Delta_A : X \rightarrow \{0, 1\}$

Si nous voulons assouplir cette subsomption stricte d'un objet au concept A, nous devons définir le concept A par un sous-ensemble flou. Ce sous-ensemble flou est défini par une fonction d'appartenance

$F_A : X \rightarrow [0, 1]$

Un élément de X tombera donc sous le concept suivant un certain degré $F_A(x)$. Si $F_A(x) = 1$ l'élément appartient entièrement au concept, et si $F_A(x) = 0$ il n'y appartient pas du tout.

Avant de développer plus largement cette théorie, il convient de s'interroger sur la précision des propriétés des objets manipulés par notre système (*i.e.* les corpus, les documents et les séquences). Nous pouvons facilement trouver des propriétés imprécises. Dans une tâche d'annotation des entités nommées il est courant d'associer la classe sémantique à laquelle appartient l'entité nommée, une association difficile à réaliser dans certains contextes. Dans l'exemple suivant

Après le 11 septembre l'Amérique est partie en guerre,

la séquence *11 septembre* est une entité nommée mais en l'absence d'information supplémentaire il est difficile de savoir s'il s'agit d'un point de départ (une date simple) ou d'une cause (la date rappelant l'attentat). Notons qu'il ne s'agit pas d'une annotation concurrente pour la valeur de la classe de l'entité nommée. Les deux valeurs *Evènement* et *Date* sont simultanément vraies. L'auteur peut vouloir entrenir cette ambiguïté dans la suite de son discours, certaines propositions n'étant alors cohérentes que si le lecteur interprète cette entité nommée comme un événement et d'autres propositions comme une date. Si le système est incapable de manipuler des informations imprécises, en raison d'une représentation qui l'oblige à choisir une interprétation, il lui sera alors impossible de conserver la cohérence du discours et de retenir les bonnes

informations dans ses inférences.

Toutefois la grande majorité des attributs dont nous nous servons pour la tâche de résolution des anaphores sont précis. On peut dire clairement si une séquence est un pronom ou si elle est précédée d'un article défini. Seuls les termes du domaine pourraient être représentés avec imprécision. Les informations imprécises sont trop peu nombreuses pour être réellement significatives dans la décision d'un système de résolution. Nous choisissons par conséquent d'ignorer leurs imprécisions dans notre représentation.

3.2 Des informations incertaines et manquantes

Pour un système d'inférence automatique, la logique classique du 1^{er} ordre serait une très bonne représentation, si la valeur de vérité des informations que ce système manipule était totalement déterminée lors des calculs. Nous distinguons les informations d'entrée qui sont l'ensemble des propriétés dont la valeur de vérité est communiquée au système (ex. le vocabulaire d'un langage), des informations produites, l'ensemble des annotations ajoutées au document (resp. au corpus) par le système (ex. les mots présents dans un document).

Les informations d'entrée manipulées par le système sont les premières informations qui sont susceptibles d'être manquantes et incertaines. Ces informations proviennent d'une annotation humaine ou automatique et sont rarement exemptes d'erreurs et d'oublis, par exemple, une entité nommée mal annotée ou oubliée. Comme ces informations appartiennent à l'ensemble des attributs utilisés lorsque notre système infère les informations produites, ces dernières héritent de la qualité des informations d'entrée. Notre système doit tenir compte de l'imperfection des informations qu'il exploite afin d'estimer la qualité de ses conclusions. Le système doit être capable de réviser ses conclusions lorsque la certitude d'un ensemble d'informations vient à être connue. De nombreux modèles d'inférences permettant la révision des conclusions ont été proposés. Ils sont traditionnellement séparés en deux grands groupes : les méthodes d'inférences symboliques et numériques. Après une rapide présentation de chaque groupe nous déterminerons notre préférence pour les méthodes numériques et défendrons notre choix pour les réseaux bayésiens.

3.3 L'inférence symbolique

La pemière possibilité pour modéliser le raisonnement d'un agent (humain ou artificiel) est la modélisation symbolique par le biais de la logique. Cependant toutes les logiques n'offrent pas à l'agent la possibilité de réviser ses conclusions. Les logiques qui présentent la propriété de monotonie sont incapables de rendre compte de la révision des connaissances d'un agent.

> *Propriété de monotonie :*
>
> soit ϕ une formule d'un langage logique, Γ, Δ un ensemble de formules définies sur le même langage et $\vdash$ une relation de déduction définie sur un système formel,
>
> la logique est monotone ssi si $\Gamma \vdash \phi$ alors $\Gamma, \Delta \vdash \phi, \Delta$.

Une conclusion qu'un agent est capable d'inférer sur un ensemble d'hypothèses sera toujours déductible même lorsque les connaissances de l'agent augmentent. Cette propriété est idoine

lorsque l'agent réfléchit sur un domaine où les connaissances sont toutes accessibles, certaines et intemporelles comme les mathématiques. Mais elle est inadaptée lorsqu'il réfléchit sur un domaine où il ne connaît qu'une partie des hypothèses nécessaires à l'inférence, où encore si les hypothèses qu'il tient pour vraies sont en réalité fausses (ou peuvent devenir fausses si le temps est un paramètre du domaine[15]). Un ensemble de logiques dites non-classiques ont été proposées pour formaliser le raisonnement sur de tels domaines.

3.3.1 Raisonnement révisable

Dans certaines situations l'agent ne possède qu'une partie des hypothèses nécessaires au raisonnement qu'il doit réaliser. C'est le cas par exemple lorsque des informations lui sont délibérement cachées ou lorsque l'observation directe d'une partie des informations est impossible. L'agent est alors contraint de déterminer un ensemble des conclusions temporaires et cohérentes avec les informations à sa disposition. Lorsqu'une information requise, jusqu'alors inconnue, devient disponible, l'agent doit reconsidérer ses conclusions en tenant compte de cette nouvelle information.

Un grand nombre de logiques ont été proposées pour permettre la révision des conclusions d'un agent. Ces logiques ont toutes la propriété de non-monotonie :

Une logique est non-monotone ssi il existe ϕ tel que $\Gamma \vdash \phi$ et $\Gamma, \Delta \nvdash \phi, \Delta$.

Pour une étude détaillée des différentes logiques dites non-monotones et de leurs relations nous renvoyons le lecteur aux synthèses suivantes [Thayse *et al.*, 1990a], [Thayse *et al.*, 1990b], [Kayser, 1997]. Parmi ces différentes logiques considérons en exemple l'une des plus connues, la logique des défauts de Reiter. Cette logique a été initialement conçue pour modéliser le raisonnement prototypique. Pour suppléer aux prémisses manquantes, l'agent va introduire dans le raisonnement des hypothèses traduisant ses connaissances prototypiques, *i.e.* des hypothèses inférées à partir d'un ensemble de faits ou de propriétés qu'il a le plus souvent observés ou qu'il sait être les plus communs. Lorsque l'agent découvre la valeur d'une prémisse manquante il doit reconsidérer ses précédentes conclusions et les réviser le cas échéant. Supposons que la théorie parle d'un oiseau mais qu'elle ne précise pas l'espèce de l'oiseau, l'agent peut conclure, *par défaut* que l'oiseau vole, car la plupart des espèces d'oiseaux volent. Si maintenant l'espèce de l'oiseau est connue et qu'il s'agit d'une autruche, l'agent doit modifier ses conclusions et conclure que l'oiseau ne vole pas.

3.3.2 Révision des hypothèses

Les logiques non-monotones précédentes ne traitent qu'un aspect de la révision des conclusions, celui des hypothèses indisponibles. Elles ne rendent pas compte de la fiabilité des informations connues par l'agent. Une information nouvellement apprise peut être fausse car provenant d'une mauvaise observation ou d'une mauvaise source, ou lorsque cette information est vraie, elle peut être inconsistante avec une hypothèse que l'agent croit être vraie alors que l'hypothèse est fausse. Les mécanismes classiques de raisonement des logiques non-monotones conduisent dans ces cas à des théories inconsistantes.

[15] Les objets que nous manipulons dans cette étude, *i.e.* les séquences qui composent un document, ne sont pas des objets évoluant dans le temps, nous ne présenterons donc pas les logiques temporelles.

Pour traiter l'arrivée d'une information inconsistante avec l'ensemble des hypothèses crues, l'agent doit être capable de remettre en question la nouvelle information ou le sous-ensemble des hypothèses qui entre en contradiction avec cette dernière. Cette remise en question se traduit par l'ajout ou la suppression d'un sous-ensemble d'hypothèses pour obtenir une nouvelle théorie consistante et compatible avec la nouvelle information. Les logiques non-monotones du paragraphe précédent permettent de déduire un nouvel ensemble de conclusions à l'arrivée d'une nouvelle information mais n'offrent aucun mécanisme de recherche des hypothèses à retirer pour retrouver une nouvelle théorie consistante. Elles sont, de ce fait, dites statiques par opposition aux logiques de raisonnement non-monotone dynamiques. Le lecteur intéressé pourra se repporter à la thèse de [Fabiani, 1996] pour une présentation de ces logiques.

3.3.3 Abandon de l'inférence symbolique pour le TAL

Poursuivons notre parallèle entre le raisonnement d'un agent et d'un système de TAL pour déterminer la logique la plus adaptée au TAL. Nous pouvons assimiler les informations d'entrée du système à l'ensemble des hypothèses tenues pour vrai par l'agent. Le système utilise ces informations d'entrée pour calculer les informations de sortie qui sont alors l'ensemble des conlusions cohérentes avec les hypothèses.

Lors des premières étapes d'annotations (ex. la segmentation en phrases ou en mots) ou lorsque le système peut éviter d'ajouter une nouvelle annotation, il est toujours possible de calculer une solution cohérente avec les informations d'entrées même si ces dernières contiennent des informations erronées. Dans le cadre d'une tâche d'annotation des entités nommées, par exemple, si l'ajout d'une entité pose problème, le système ne fait rien et conserve ainsi la cohérence de la théorie. La solution calculée est cohérente mais certainement fausse et elle devient une nouvelle information d'entrée erronée pour les tâches d'annotation suivantes. Pour d'autres tâches le système doit produire une annotation et lorsqu'il a trop d'informations erronées en entrée, il lui est impossible de calculer une solution (*i.e.* déduire une conclusion cohérente avec l'ensemble des hypothèses). Ainsi lors d'une analyse syntaxique en constituants, le système doit produire un arbre syntaxique pour une phrase en s'appuyant sur un étiquetage des mots. La présence de mots inconnus ou mal étiquetés rendent l'analyse de certaines partie de la phrase impossible. Si le système ne peut modifier l'étiquette de certains mots, il n'a d'autre solution que de laisser ces parties de phrases non analysées.

Cette dernière situation correspond exactement au cas d'application d'une logique non-monotone dynamique. Le TAL est un processus incrémental où les informations produites lors d'une tâche d'annotation sont employées, sans aucune garantie de véracité, comme informations d'entrée pour une autre tâche. Lorsqu'une tâche devient impossible à réaliser, il faut alors rechercher puis corriger les valeurs erronées des informations d'entrée.

Les travaux récents ci-dessous mettent en évidence la nécessité d'employer une logique non-monotonne pour réaliser une tâche d'annotation, mais ignorant la présence d'informations d'entrée érronées, ils ne proposent que des systèmes raisonnant à partir de logiques non-monotonnes statiques : la logique des défauts avec priorité[16] pour les travaux que nous citons.

[16]La logique des défauts classique substitue aux informations manquantes des informations prototypiques pour permettre de calculer un ensemble de conclusions cohérentes. Seulement dans certains cas le choix d'une information prototypique au profit d'une autre information prototypique conduit au calcul de deux conclusions incompatibles. Or il n'y a aucune raison de préférer une information plutôt qu'une autre. Pour se prononcer en faveur d'une

Dans [Philippe *et al.*, 2003] les auteurs rapprochent la *Théorie Optimale* avec la logique des défauts avec priorité. La Théorie Optimale est une méthode qui permet d'exprimer un ensemble de contraintes grammaticales pour sélectionner le meilleur candidat possible. L'expression des contraintes dans la langue naturelle nous libère de la représentation des structures des candidats et permet d'appliquer cette méthode à un grand nombre de tâches du TAL, en particulier la résolution des anaphores, voir [Byron & Gegg-Harrison, 2004]. Les contraintes grammaticales sont constituées par un ensemble de règles ordonnées qui peuvent être violées. Une similarité avec les caractéristiques des règles de défauts autorise la reformulation de ces contraintes par les règles de défauts avec priorités et rend possible l'implémentation logique d'une théorie particulière. Les recherches préliminaires de [Cohen, 2007] poursuivent le travail précédent pour la résolution des anaphores. L'auteur choisit de représenter le sens du discours au moyen des DRS. Dans ce cadre logique la relation d'antécédance est une simple relation d'identité. Toute nouvelle entité est identifiée aux entités précédemment introduites dans le discours tant que le système ne peut prouver le contraire. Pour apporter la preuve de leurs différences, le système doit appliquer une série de défauts selon leur ordre de priorité. Ces règles de défauts instancient les facteurs utilisés habituellement pour la résolution des anaphores. Cette approche est encore récente et aucune traduction des facteurs en règles de défauts n'a été proposée.

Les inférences des logiques non-monotones dynamiques semblent être une solution possible pour traiter les informations incertaines et manquantes du TAL. Ces formalismes héritent des avantages des méthodes symboliques. Les règles, exprimables en logique du première ordre, sont explicites et peuvent être comprises rapidement par un utilisateur averti. Les décisions et les comportements du système peuvent toujours être anticipés et expliqués.

Toutefois plusieurs raisons nous ont poussé à préférer le cadre du raisonnement numérique pour cette étude. En premier lieu, la description complète des règles nécessaires à la résolution des anaphores requiert un gros travail de modélisation dont le chantier vient juste de commencer. Ensuite, les mécanismes de révision des hypothèses en cas de conclusions inconsistantes, qui sont nécessaires pour notre domaine compte tenu de la qualité des informations d'entrée des systèmes, sont complexes et encore imparfaitement maîtrisés [Hansson, 2003]. Ils ne garantissent pas toujours le résultat intuitif attendu. Enfin, l'obtention d'une conclusion inconsistante est une conséquence possible lorsque l'agent tient pour vrai un certain nombre d'hypothèses. Une solution plus simple que de doter l'agent d'un mécanisme de révision des croyances consiste à le faire douter de toutes les connaissances qu'il manipule. C'est la solution que nous avons choisie et mise en œuvre grâce aux probabilités et que nous présentons dans la section suivante.

conclusion, il faut étendre le formalisme de la logique des défauts en introduisant des priorités entre les informations prototypiques, comme le propose [Brewka & Eiter, 2000]. La conclusion calculée au moyen de l'information prototypique prioritaire sera préférée à l'autre conclusion.

3.4 L'inférence numérique

3.4.1 Les théories pour la représentation de l'incertitude

La notion d'information incertaine

Avant d'analyser les différentes représentations de l'incertitude pour en choisir une, examinons la signification de ce concept. Lorsque nous fixons la valeur d'un attribut d'un objet et plus généralement lorsque nous affirmons qu'un énoncé d'un langage logique est vrai, nous décrivons un ensemble de mondes possibles et nous rejetons les autres mondes (les mondes qui rendent faux l'énoncé). Lorsque tous les attributs de tous les objets évoluant dans notre monde sont fixés, nous décrivons un monde unique. Maintenant supposons que notre agent soit incertain de la valeur d'un attribut, il ne peut restreindre le nombre de mondes possibles qu'il considère dans son inférence. Plus le nombre d'attributs dont l'agent doute est grand, plus le nombre de mondes possibles qu'il doit envisager est important et d'autant moins fiables seront ses conclusions. Dans certaines situations, bien que l'agent soit incertain de l'information, il peut déterminer un certain degré de confiance dans sa véracité et préférer un sous-ensemble de mondes possibles. S'il ne mesure pas son incertitude, il perd une information nécessaire lors de l'inférence et peut être conduit à des conclusions erronées.

Supposons que notre agent doive deviner le résultat d'un tirage de deux dés, sans autre information, l'agent ne doit exclure aucun des 36 résultats possibles. Si on lui apporte une information supplémentaire, par exemple que le résultat du premier dé est égal à 1, il peut ignorer 30 des résultats possibles et ne considérer que six mondes possibles. Notons que dans ce cas de figure, il n'y a aucune raison pour l'agent de préférer un monde particulier parmi les six mondes possibles. Maintenant supposons que les deux dés soient pipés et tombent, de fait, plus souvent sur six ou sur cinq. Comme précédemment il sait que le résultat du premier dé est 1. Il peut donc se restreindre à six mondes possibles, mais cette fois il peut préférer deux mondes parmi les six restants en raison du biais du dé. Notons que c'est une préférence et non une certitude, les autres mondes sont toujours possibles.

Un grand nombre de travaux proposent différentes représentations de l'incertitude [Halpern, 2003]. L'étude de chacune de ces représentations, qui demanderait de les opposer ou de les unifier, sort du cadre de ce travail. Il s'agit plutôt pour nous de comparer leur pouvoir d'expression et d'établir leurs conditions d'application en vue de choisir la plus adaptée à notre problème de TAL.

Probabilités, possibilités et évidences

Probabilités : Les objets et les attributs avec lesquels nous travaillons sont toujours en nombre fini, même si ce nombre peut être très grand[17]. Nous définissons X un ensemble de mondes possibles fini et l'ensemble des parties de X, noté $\mathfrak{P}(X)$ tel que :

$y \in \mathfrak{P}(X)$ ssi $y \subseteq X$

La certitude d'un agent peut être perçue comme un poids subjectif que ce dernier répartit sur un monde possible ou un ensemble de mondes possibles. La première et la plus connue des

[17] Toutefois les théories ci dessous peuvent être étendues pour travailler avec un ensemble de mondes possibles infini, voir [Halpern, 2003].

représentations de l'incertitude est celle des probabilités. On définit traditionnellement une mesure de probabilité comme une fonction définie sur l'ensemble des parties des mondes possibles et à valeur dans l'intervalle [0,1]. Cette fonction doit respecter les contraintes suivantes :

$Pr : \mathfrak{P}(X) \rightarrow [0, 1]$

– $Pr(X) = 1$

– $Pr(\varnothing) = 0$

– $\forall A, B \in \mathfrak{P}(X)$ si $A \cap B = \varnothing$, $Pr(A \cup B) = Pr(A) + Pr(B)$

Un sous-ensemble de mondes qui reçoit une probabilité égale à 0 est jugé impossible et nécessaire (ou vrai) si la probabilité est égale à 1.

On peut aussi définir une distribution de probabilité à partir de la mesure de probabilité. Soit $A_1, A_2, ..., A_n$ une famille d'ensembles mutuellement exclusifs et définis sur $\mathfrak{P}(X)$ tels que Pr($\bigcup_{i=1,..,n} A_i$)=1 :

$p : A_1, A_2, ..., A_n \rightarrow [0, 1]$

$p(A_i)=Pr(A_i)$ pour $1 \leq i \leq n$

La question de l'interprétation d'une mesure de probabilité est une question philosophique toujours débattue aujourd'hui. Sans aucune prétention polémique, nous choisissons d'introduire la mesure de probabilité comme une mesure subjective de la certitude d'un agent sur la véracité d'un ensemble de mondes possibles. Cette interprétation subjective des probabilités est suffisamment libre pour autoriser la définition d'une mesure de probabilité quelle que soit l'origine de cette croyance.

Le cadre d'application des probabilités est le cadre d'application où l'agent est le plus informé de son environnement. Il peut recourir à différentes sources d'informations telles que les résultats statistiques d'un ensemble d'expériences similaires, l'observation directe ou indirecte d'instruments ou encore l'évaluation d'un expert traduite numériquement pour construire un modèle *a priori* complet de cet environnement.

La représentation probabiliste est la mesure la plus précise de l'incertitude : étant donnée une mesure de probabilité on peut toujours calculer une distribution de probabilité. Chaque monde possible reçoit ainsi un poids correspondant au degré de croyance de l'agent pour que ce monde soit le monde réel. Toutefois la définition d'une mesure de probabilité peut ne pas toujours être complète. Si l'agent est capable d'assigner une probabilité à un sous-ensemble de mondes, il peut rester sans avis sur les sous-ensembles restants, ou ne pas vouloir être précis sur la valeur d'une probabilité.

Supposons que notre agent s'intéresse au résultat d'une course de chevaux. La course compte 4 chevaux $\{a, b, c, d\}$ au départ. Notre agent a appris que le cheval a est malade, il fixe donc la probabilité du singleton $Pr(\{a\}) = 0$ mais sans autre information, il est incapable de déterminer les probabilités des autres sous-ensembles. Comme le cheval a ne peut gagner et que nous supposons que la course doit avoir un gagnant, nous savons que $Pr(\{b, c, d\}) = 1$ mais il est impossible d'assigner une probabilité aux singletons de cet ensemble. Une méthode classique consiste alors à répartir de manière égale la probabilité entre tous les mondes possibles :

soit $Pr(\{b\}) = 1/3, Pr(\{c\}) = 1/3$ et $Pr(\{d\}) = 1/3$

Cette distribution uniforme de la probabilité marque l'ignorance de l'agent, puisque ce dernier ne peut préférer aucun monde possible.

Une première solution, qui permet de rester dans le cadre de la théorie des probabilités, consiste à choisir une mesure de probabilité et d'accompagner ce choix d'une probabilité traduisant l'incertitude sur cette probabilité. Il s'agit alors de se doter d'une probabilité de second

ordre, une question qui est discutée dans [Fabiani, 1996][18].

Possibilités et Nécessités : Une seconde solution est de changer de mesure. Dans le cadre de la théorie des possibilités, les connaissances de l'agent sont insuffisantes pour qu'il puisse établir une mesure de probabibilité raisonnable. Il va alors se servir de deux mesures complémentaires de l'incertitude pour raisonner.

La mesure de possibilité est une estimation réalisée par l'agent de la possibilité que le monde réel soit contenu dans un sous-ensemble donné, elle se calcule grâce à une fonction Π :

$\Pi : \mathfrak{P}(X) \rightarrow [0, 1]$

$\Pi(X) = 1, \Pi(\varnothing) = 0$, et, pour toute famille d'ensembles de $\mathfrak{P}(X)$,

$\Pi(\cup_{i=1,2,...} A_i) = max_{i=1,2,...} \Pi(A_i)$

Seule, cette mesure ne dit rien de la nécessité que le monde réel soit dans un sous-ensemble de mondes possibles. La définition de Π n'interdit pas le cas où

$\Pi(A) = 1$ et $\Pi(A^c) = 1$, où A^c est le sous-ensemble complémentaire de A.

Or la signification de $\Pi(A) = 1$ est très différente suivant que $\Pi(A^c) = 0$ ou que $\Pi(A^c) = 1$. Dans le premier cas il est non seulement possible mais aussi nécessaire que A contienne le monde réel alors que dans le second cas nous ne savons rien de la "position" du monde réel car s'il est possible qu'il soit contenu dans A, il est aussi possible qu'il soit dans son complémentaire.

Pour rendre compte de la nécessité d'un sous-ensemble A il faut compléter la mesure de possibilité par une mesure de nécessité :

$N : \mathfrak{P}(X) \rightarrow [0, 1]$

$N(\varnothing) = 0, N(X) = 1$, et pour toute famille d'ensembles de $\mathfrak{P}(X)$,

$\Pi(\cap_{i=1,2,...} A_i) = min_{i=1,2,...} \Pi(A_i)$

La propriété suivante permet de lier ces deux notions duales :

$\forall A \in \mathfrak{P}(X) N(A) = 1 - \Pi(A^c)$

Un sous-ensemble contiendra d'autant plus nécessairement le monde réel que son complémentaire est impossible.

La donnée des deux mesures de possibilité et de nécessité définit l'encadrement d'un ensemble de mesures de probabilités admissibles, si cet ensemble existe, sur l'ensemble des sous-ensembles de mondes possibles (voir [Fabiani, 1996] pour une justification formelle) :

$\forall A \in \mathfrak{P}(X), N(A) \leq Pr(A) \leq \Pi(A)$

Cet encadrement peut être précisé selon les différents degrés de croyance de l'agent :

- Lorsque l'agent est certain que le monde réel est dans un sous-ensemble A, toutes les mesures de probabilité possibles marquent cette certitude par une probabilité unité, $Pr(A) = 1$, ce qui se traduit dans le cadre de la théorie des possibilités par une nécessité unité : N(A)=1 et on obtient alors $1 = N(A) \leq Pr(A) = 1$
- Lorsque l'agent sait qu'un sous-ensemble A de mondes est impossible, il lui assigne une probabilité nulle, $Pr(A) = 0$, de même qu'une mesure de possibilité nulle $\Pi(A) = 0$ et on obtient $0 = Pr(A) \leq \Pi(A) = 0$

[18] Nous proposons cette solution dans les perspectives de la section 7.2.3 car elle est facile à mettre en application dans le cadre des réseaux bayésiens.

- Enfin si l'agent estime un sous-ensemble de mondes possible, il lui assignera une probabilité positive ce qui peut être traduit de 3 manières différentes dans le cadre de la théorie des possibilités :
  - l'agent pense que ce sous-ensemble n'est pas tout à fait possible et donc non nécessaire soit $\Pi(A) \lneq 1$, $N(A) = 0$ et $\Pi(A^c) = 1$
    Dans ce cas toute mesure de probabilité admissible est encadrée comme suit :
    $0 \leq P(A) \leq \Pi(A) \lneq 1$ et $0 \lneq 1 - \Pi(A) \leq Pr(A^c) \leq 1$
  - l'agent pense que le sous-ensemble est complètement possible et même un peu certain, $N(A) \gneq 0$, $\Pi(A) = 1$ et $\Pi(A^c) = 0$
    Les mesures admissibles sont alors encadrées ainsi :
    $0 \lneq N(A) \leq Pr(A) \leq 1$ et $0 \leq Pr(A^c) \leq 1 - N(A) \lneq 1$
  - l'agent ignore la nécessité du sous-ensemble mais le pense complétement possible (cas d'ignorance dans cette théorie), $\Pi(A) = 1$ et $N(A) = 0$
    et l'encadrement des mesures de probabilité est le suivant :
    $0 = N(A) \leq Pr(A) \leq \Pi(A) = 1$ et $0 = N(A^c) \leq Pr(A^c) \leq \Pi(A^c) = 1$

En l'absence d'information numérique fiable pour guider son jugement, l'agent doit se reposer sur une évaluation subjective de l'incertitude. Une évaluation subjective reste difficile à quantifier précisément. Il est donc appréciable de pouvoir définir l'incertitude avec plus de souplesse grâce à un encadrement donné par la possibilité et la nécessité. Cependant lorsqu'aucune évaluation numérique fiable ne vient étayer l'évaluation subjective d'un agent, on est en droit d'interroger l'exactitude de cette évaluation. Pour se doter d'une évaluation plus fiable, l'agent peut comparer des évaluations concurrentes en vue de formuler, lorsque c'est possible, une nouvelle évaluation conciliant les précédentes évaluations concurrentes. La théorie des possibilités n'offre pas un calcul simple pour réaliser ce compromis et une autre théorie est en général utilisée dans ce but, la théorie de l'évidence de Dempster-Shafer.

Croyance et Plausibilité : Nous présentons ici une théorie de l'évidence simplifiée [Bouchon-Meunier, 1995]. La degré de confiance employé par l'agent pour raisonner sur l'incertitude est, dans cette théorie, interprété comme une masse de croyance que l'agent répartit uniquement sur les sous-ensembles des mondes pour lesquels l'agent pense qu'ils contiennent le monde réel. La masse répartie sur ces sous-ensembles est ensuite employée pour déterminer une borne inférieure et une borne supérieure d'incertitude pour l'ensemble de mondes possibles.

On définit une fonction m qui associe une masse de croyance à chaque sous-ensemble de l'ensemble des mondes possibles. Une masse de croyance est un coefficient compris entre 0 et 1 :

$m : \mathfrak{P}(X) \rightarrow [0, 1]$

La répartition des masses de croyance doit respecter les deux contraintes suivantes :

$\Sigma_{A \in \mathfrak{P}(X)} m(A) = 1$ et m($\varnothing$)=0

On appelle *élément focal* tout sous-ensemble E non vide de X tel que $m(E) \neq 0$, et on note **E** leur ensemble.

E est donc l'ensemble des mondes que l'agent croit, même un petit peu, possibles. La confiance minimale que nous avons dans un sous-ensemble de mondes possibles quelconque est la somme de tous les éléments focaux appartenant au sous-ensemble. Elle se traduit par la fonction de croyance :

$Bel : \mathfrak{P}(X) \rightarrow [0,1]$
$Bel(A) = \Sigma_{E \in \mathbf{E}, E \subseteq A} m(A)$

La croyance maximale, calculée par la fonction de plausibilité, est obtenue en ajoutant aux masses des éléments focaux de notre sous-ensemble les masses des éléments de tout autre ensemble lié à ce dernier. Deux sous-ensembles sont *liés* lorsqu'ils partagent au moins un monde en commun. La plausibilité se définit formellement comme suit :

$Pl : \mathfrak{P}(X) \rightarrow [0,1]$
$Pl(A) = \Sigma_{E \in \mathbf{E}, E \cap A \neq \varnothing} m(A)$

Dans le cas particulier où l'agent est incapable de préférer un sous-ensemble de mondes au profit d'un autre sa masse de croyance est répartie entièrement sur X et on obtient pour tout sous-ensemble $A \neq$X

Bel(A)=Bel(A^c)=0 et *Pl(A)=Pl(A^c)=1*

La théorie de l'évidence est la théorie la plus permissive dans la mesure où elle permet de croire à n'importe quels sous-ensembles de mondes possibles, qu'ils soient cohérents ou non entre eux. Les valeurs définies par les fonctions de croyance et de plausibilité encadrent toutes les valeurs possibles d'une mesure de probabilité compatible avec la distribution de la masse de croyance (pour plus de détail, voir [Bouchon-Meunier, 1995]).

Le principal intérêt de cette théorie est l'existence d'une loi simple pour composer une nouvelle distribution de masses de croyance à partir de deux distributions différentes :

Soit A un sous-ensemble de mondes possibles de X, il est possible de calculer une distribution de masse de croyance m_{12} à partir de deux distributions m_1 et m_2 s'il existe au moins deux sous-ensembles de X, B et C, tel que $B \cap C \neq \emptyset$ et $m_1(B).m_2(C) \gneq 0$.

Formellement, on définit $m_{12}(\emptyset) = 0$ et $m_{12}(A) = k\Sigma_{B \subseteq X, C \subseteq X, B \cap C = A}$ où k un coefficient normalisateur.

Ce cadre théorique est le cadre le plus adapté lorsque les informations à disposition de l'agent sont très faibles. Ne disposant pas d'une évaluation numérique fiable pour justifier une mesure de probabilité, l'agent doit s'en remettre à une évaluation subjective de l'incertitude. La qualité de l'évaluation formulée par un seul expert repose sur la fiabilité de cet expert. Pour en garantir la qualité, une solution consiste à formuler l'évaluation à partir de différentes expertises. En comprenant le degré de certitude comme un simple poids de croyance, il devient possible de combiner facilement des expertises différentes si les avis exprimés ne sont pas complémentement opposés (*i.e.* s'il existe au moins un sous-ensemble de mondes possible commun dans chaque couple de distribution de masse de croyance à agréger, soit l'intersection de B et C dans notre définition).

3.4.2 Le choix du cadre théorique des probabilités

Pour choisir le cadre théorique avec lequel nous allons représenter l'incertitude de notre système, nous devons identifier les sources d'informations à notre disposition pour mesurer l'incertitude des annotations à partir desquelles nous travaillons en TAL.

Un corpus annoté est la première source d'informations, et généralement la source principale en TAL, que nous pouvons exploiter. Les fréquences des phénomènes linguistiques observés en corpus, lors des étapes d'apprentissage des systèmes, fournissent une base statistique que nous pouvons exploiter pour fixer des mesures de probabilité fiables. Nous présentons des exemples

concrets pour l'apprentissage d'une mesure de probabilité à partir de données annotées dans les sections 4.1.3 et 5.1.3.

En l'absence de corpus annotés, une seconde source d'informations peut servir pour établir une mesure de probabilité ou seulement la compléter : l'expertise. L'interrogation d'un linguiste est bien sûr l'expertise la plus naturelle pour un problème de TAL, mais le recours à d'autres expertises moins subjectives sont envisageables. Supposons que l'on cherche à évaluer la certitude d'une annotation des entités nommées sans avoir de corpus annoté disponible. L'extrapolation d'une mesure de probabilité à partir des performances publiées de l'outil utilisé pour cette annotation ou encore l'extrapolation à partir d'un échantillon créé à cette fin fournissent souvent de bons résultats.

Ainsi nous pouvons dans cette étude faire l'hypothèse que les informations que nous avons à disposition sont suffisantes pour déterminer complétement un modèle *a priori* que l'agent interrogera pour se guider lors de son raisonnement. Nous travaillerons donc avec la théorie des probabilités.

3.4.3 Raisonner sur des informations probabilisées

Pour raisonner à partir d'informations probabilisées, nous avons besoin d'introduire quelques notions classiques de la théorie des probabilités.

On appelle *variable aléatoire discrète* (VA) une fonction de l'ensemble des sous-ensembles de mondes possibles $\mathfrak{P}(X)$ vers un ensemble fini quelconque :

$A : \mathfrak{P}(X) \rightarrow \{a_1, ..., a_n\}$

Une VA est un simple nom pour pouvoir parler d'un sous-ensemble de mondes possibles que l'on veut distinguer. L'expression $\{A = a_i\}$ dénote alors le sous-ensemble de mondes $\{x/A(x) = a_i\}$ où A est une VA et a_i appartient au co-domaine de A.

Les attributs peuvent être vus comme des VA et nous pouvons définir une distribution de probabilités sur l'ensemble des valeurs prises par un attribut[19].

Lorsque nous fixons les valeurs des VA $\{A = a_i\}$ et $\{B = b_i\}$, nous désignons, par définition, deux ensembles de mondes de $\mathfrak{P}(X)$. Il est donc possible de réaliser l'intersection de deux VA et de lui assigner une probabilité, la probabilité associée à l'intersection de ces deux ensembles. Cette probabilité est définie par une distribution de probabilité que l'on appelle distribution de probabilité jointe[20]. Soit A, B, C trois VA définies sur un même ensemble de mondes possibles :

$Pr_{ABC} : \{a_1, ..., a_m\} \times \{b_1, ..., b_n\} \times \{c_1, ..., c_p\} \rightarrow [0, 1]$

tels que $Pr_{ABC}(a_i, b_j, c_k) = Pr(A = a_i \cap B = b_j \cap C = c_k)$, où $i \in [1, m], j \in [1, n], k \in [1, p]$.

Toutefois, bien que connaissant les probabilités que les deux VA A et B prennent respectivement les valeurs a_i et b_j, il n'existe pas de fonction générale pour déterminer leur probabilité jointe $P(A = a_i \cap B = b_j)$. Cette dernière dépend des liens d'influence qu'entretiennent les 2 VA entre elles. Lorsqu'elles sont indépendantes nous savons que :

[19] Les sous-ensembles de mondes dénotés par les différentes valeurs d'un attribut sont mutuellement exclusifs et exhaustifs (i.e. si les valeurs possibles pour un attribut sont $a_1, ..., a_n$ alors quelque soit $x \in X$ où X est notre ensemble de mondes de référence il existe a_i ($1 \leq i \leq n$) vrai pour x).

[20] On peut étendre cette définition de manière identique à un nombre quelconque fini de VA.

$P(A = a_i \cap B = b_j)$=$P(A = a_i).P(B = b_j)$
mais lorsqu'elles ne le sont pas, nous ne pouvons rien dire sans connaître ce lien d'influence.

Dans la théorie des probabilités, le lien d'influence entre plusieurs VA est fourni par la notion de probabilité conditionnelle. On note $P(A = a_i|B = b_j \cap C = c_k)$ la probabilité de A en l'état a_i sachant que B et C sont en l'état b_j et c_k ce qui doit se comprendre comme la probabilité d'observer $A = a_i$ lorsqu'on suppose que $B = b_j$ et $C = c_k$ sont vrais. On calcule cette probabilité grâce à l'équation suivante :
$P(A = a_i|B = b_j \cap C = c_k)$=$\frac{P(A=a_i \cap B=b_j \cap C=c_k)}{P(B=b_j \cap C=c_k)}$
Une fois connu, cet ensemble des distributions de probabilité, le système est maintenant en mesure de les utiliser pour raisonner et déterminer les nouvelles probabilités d'un ensemble de mondes possibles lorsque de nouvelles connaissances lui sont connues. Il existe plusieurs méthodes possibles de raisonnement. Le choix varie en fonction du nombre de VA d'intérêts, des relations d'influence entre les VA du modèle, des connaissances observables, de la confiance accordée aux observations, *etc*. Le chapitre suivant présente les modèles les plus connus et justifie notre préférence pour le modèle des réseaux bayésiens.

3.5 Conclusion

Dans ce chapitre nous avons caractérisé l'imperfection des annotations utilisées et produites par les systèmes de TAL. Cette imperfection est une notion que nous définissons comme la conjonction inclusive de trois concepts distincts : l'imprécision d'une annotation, son incertitude ou encore son absence. Bien que les annotations que nous utilisons en TAL peuvent être imprécises pour la tâche particulière de la résolution des anaphores pronominales, le nombre d'annotations imprécises que nous pouvons raisonnablement utiliser est insuffisant. Nous choisissons donc de simplifier le problème en ne représentant que l'incertitude et l'absence des annotations ou de la valeur d'un attribut.

De nombreux formalismes d'inférence ont été proposés pour traiter des informations incertaines et manquantes. On regroupe habituellement ces formalismes en deux grandes catégories, les formalismes symboliques et numériques. Au travers de notre étude des formalismes symboliques nous avons mis en évidence que seule une logique non-monotone dynamique semble adaptée pour raisonner sur les annotations du TAL : l'expression de connaissances prototypiques peuvent suppléer aux informations d'entrée manquantes et le mécanisme de révision des hypothèses corriger les informations d'entrée incertaines. Néanmoins le recours à ce formalisme pour la résolution des anaphores est une idée récente. Devant l'important travail d'écriture des règles que nous pressentons et la complexité des mécanismes de révision des hypothèses qui sont encore mal maîtrisés, nous lui préférons un formalisme numérique. Dans ce type de formalisme nous ne recherchons plus la véracité d'une information mais son degré de confiance, ce qui facilite l'expression des informations d'entrée prototypiques et leur révision (voir le chapitre 4).

Pour choisir un formalisme numérique particulier, nous avons dû évaluer le degré de confiance des informations d'entrée sur lesquelles reposent un système de TAL. Nous avons avancé l'hypothèse que l'étude d'un corpus d'apprentissage ou d'une expertise nous donne suffisamment d'informations pour déterminer avec précision notre confiance sur les informations d'entrée. Cette hypothèse posée, nous nous dotons de la mesure la plus précise de l'incertitude et déci-

dons de travailler dans le cadre de la théorie des probabilités. Dans le chapitre suivant, nous présenterons la famille des modèles d'inférences probabilistes, les modèles génératifs, et nous justifierons notre choix du modèle des réseaux bayésiens.

Chapitre 4

Les modèles génératifs

Sommaire

Dans la section précédente nous avions arrêté notre choix sur une représentation probabiliste de l'incertitude. Les modèles inventés pour raisonner à partir d'informations probabilisées sont regroupés sous la famille des modèles génératifs. Ces modèles sont dit génératifs car ils peuvent être utilisés pour engendrer des échantillons de données qui suivent la loi jointe définie par le modèle. Dans cette section nous présentons les modèles génératifs qui nous intéressent dans la suite de ce travail. Nous donnons en fin de section un ensemble d'applications de TAL utilisant ces modèles.

4.1 Les Réseaux Bayésiens

4.1.1 Définition

Compréhension intuitive des réseaux bayésiens

Le modèle des réseaux bayésien (RB) repose sur l'idée assez naturelle de représenter graphiquement les relations d'influences entre les VA d'un modèle. Un arc unit deux VA lorsqu'il existe une relation d'influence entre elles et l'absence d'arc marque leur indépendance. Le sens de l'arc précise le sens de l'influence que nous souhaitons modéliser. Est-ce que nous pensons que A influence B ou l'inverse ? L'adjonction au graphe des distributions des probabilités mesure la "force" de cette influence. Un réseau bayésien nous donne une description qualitative des relations d'influence entre les VA données par un graphe dirigé sans circuit ainsi qu'une description quantitative de ces relations au moyen d'un ensemble de distributions de probabilités conditionnelles.

Définition formelle

Pour définir un réseau bayésien nous avons besoin d'introduire la notion de graphe orienté sans circuit.

Un graphe orienté G est défini par la donnée d'un couple (s,a) où

– s est un ensemble fini $\{s_1, ..., s_2\}$ d'élements nommés sommets

– $a \subsetneq s \times s$ un ensemble de paires ordonnées dont les éléments sont nommés arcs

On appelle chemin d'un graphe G=(s,a) une suite finie d'arcs consécutifs, soit
$\{(s_{i_1}, s_{i_2}), (s_{i_2}, s_{i_3}), ..., (s_{i_{n-2}}, s_{i_{n-1}}), (s_{i_{n-1}}, s_{i_n})\} \subsetneq a$

Un graphe orienté G=(s,a) est dit sans circuit s'il n'existe pas de chemin fermé *i.e.* de chemin tel que $\{(s_{i_1}, s_{i_2}), (s_{i_2}, s_{i_3}), (s_{i_{n-2}}, s_{i_{n-1}}), (s_{i_{n-1}}, s_{i_n})\}$ et $s_{i_1} = s_{i_n}$

s_i est un parent de s_j si $(s_i, s_j) \in a$, on définit la fonction
$Parent : s \rightarrow \mathfrak{P}(s)$
$Parent(s_x) = \{s_i, ..., s_j\}$ l'ensemble des parents de s_x

Un réseau bayésien RB est défini par le couple RB=(G,P) où

– G est un graphe orienté sans circuit défini sur un ensemble de VA fini $X = \{x_1, ..., x_n\}$

– P est un ensemble de probabilités conditionnelles associé aux VA de X tel que
$P = \{p_{x_i} / p_{x_i} = (x_i | Parent(x_i))\}$ tel que $1 \leq i \leq n$

La principale propriété d'un RB repose sur la représentation graphique des liens d'indépendance des VA. Nous pouvons simplifier la loi jointe du modèle en ne considérant que les probabilités conditionnelles réèllement pertinentes dans notre modèle, soit pour une VA donnée, ne considérer que l'ensemble des VA qui l'influencent, *i.e.* l'ensemble de ses parents.

Le RB est une représentation simplifié de la loi jointe :
$P(X_1, ..., X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)...P(X_n|X_1...X_{n-1}) =$
$\Pi_{i=1}^{n} P(X_i|Parent(X_i))$

Exemple d'un réseau bayésien

L'explication intuitive sur un exemple classique aidera à comprendre plus facilement la définition et la propriété qui précèdent. Paul est chef de projet d'une petite entreprise d'informa-

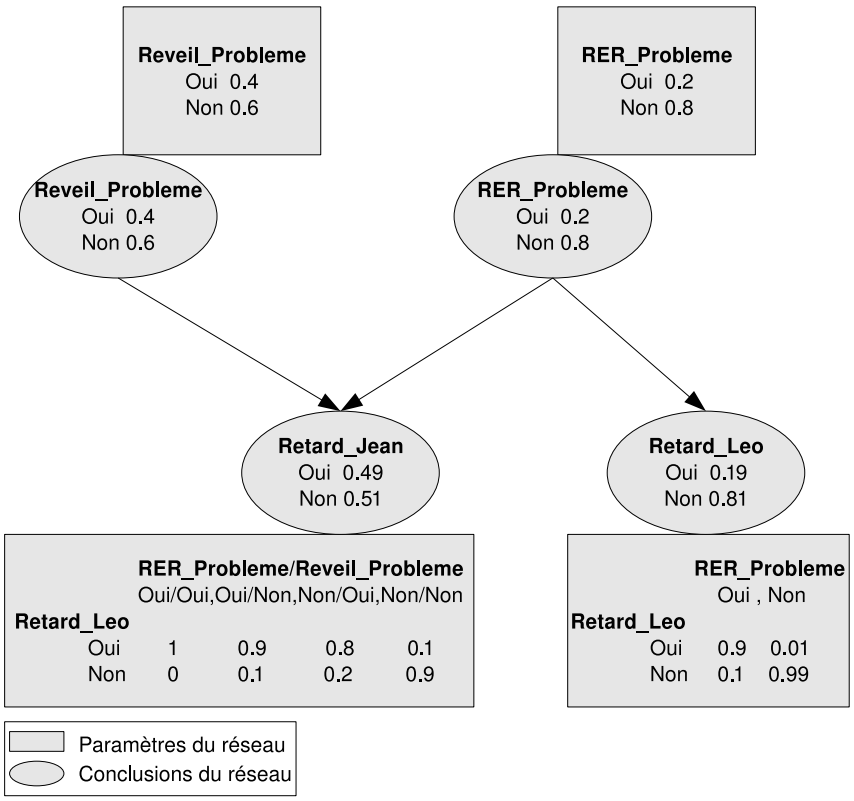

FIG. 4.1 – Exemple d'un réseau bayésien

tique. Notre chef de projet emploie deux informaticiens Jean et Léo. Jean a la fâcheuse habitude de ne pas entendre son réveil et d'arriver en retard. Léo, lui, est d'une ponctualité irréprochable, mais lui comme Jean, se rendent à l'entreprise au moyen du même RER dont les disfonctionnements sont fréquents. En cas de problème sur ce RER les deux informaticiens arrivent, sauf exception, en retard. Le graphe de la figure 4.1 décrit les relations d'influence de notre exemple. Les paramètres du réseau mesurent les connaissances *a priori* de notre chef de projet. Pendant leur collaboration Jean est souvent arrivé en retard en raison d'un problème de réveil et il décide d'ignorer les raisons qui expliquent ces problèmes de réveil. La VA *P(Réveil_Problème)* n'a donc aucun parent et il chiffre la probabilité que Jean ait un problème de réveil à 0.4, soit

P(Réveil_Problème=oui)=0.4 (et donc *P(Réveil_Problème=non)=1-0.4*).

L'arc entre les VA *RER_Problème* et *Retard_Léo* modélise l'influence des problèmes de RER sur les retards de Léo et la table de probabilité associée à la VA *Retard_Léo* établit que la probabilité que Léo arrive en retard lorsqu'il y a un dysfonctionnement sur la ligne de RER est de 0.9. Les autres VA et valeurs des tables de probabilités du réseau se comprennent de façon similaire.

La loi jointe de notre problème est décrite par l'équation[21] :

[21] Pour simplifier l'écriture de la probabilité jointe nous employons la convention d'écriture où *P(Reveil_Probleme,RER_Probleme,Retard_Jean,Retard_Leo)* résume *P(Reveil_Probleme=x∩RER_Probleme=y∩Retard_Jean=z∩Retard_Leo=w)* et se lit comme la probilité que la VA *Reveil_Probleme* prenne la valeur *x* et que la VA *RER_Probleme* prenne la

P(Reveil_Probleme, RER_Probleme, Retard_Jean, Retard_Leo)=
P(Reveil_Probleme).P(Reveil_Probleme|RER_Probleme).P(Retard_Jean|Reveil_Probleme, RER_Probleme).P(Retard_Leo|Retard_Jean,Reveil_Probleme,RER_Probleme)

La loi jointe peut être simplifiée en ne tenant compte que des liens d'influences du graphe :

P(Reveil_Probleme, RER_Probleme, Retard_Jean, Retard_Leo)=
P(Reveil_Probleme).P(RER_Probleme|Reveil_Probleme).P(Retard_Jean|Reveil_Probleme, RER_Probleme).P(Retard_Leo|Reveil_Probleme,RER_Probleme,Retard_Jean)=
P(Reveil_Probleme).P(RER_Probleme).P(Retard_Jean|Reveil_Probleme,RER_Probleme).
P(Retard_Leo|RER_Probleme)

4.1.2 L'inférence dans un réseau bayésien

L'inférence à partir d'observations certaines

Une fois le réseau défini et paramétré, il peut être employé pour inférer un ensemble de conclusions possibles qui découlent des connaissances que l'on a d'une situation. Ces conclusions sont données sous la forme de probabilités conditionnelles. Par exemple, le calcul de $P(Retard\_Jean = Non)$ détermine la probabilité de Jean d'arriver à l'heure en temps normal (*i.e.* sans aucune autre information que la connaissance de ses problèmes de réveil et la fréquence des retards du RER *a priori*), alors que le calcul de $P(retard\_Jean = Non|RER\_Probleme = Oui)$ détermine la probabilité de Jean d'arriver à l'heure lorsqu'un retard du RER a été observé (cette probabilité n'est ni égale à $P(retard\_Jean = Non|RER\_Probleme = Oui, Reveil\_Probleme = Oui) = 0$ ni égale à $P(retard\_Jean = Non|RER\_Probleme = Oui, Reveil\_Probleme = Non) = 0.1$ car aucune information concernant un problème de réveil n'est disponible).

Pour calculer la valeur exacte d'une probabilité conditionnelle quelconque du réseau nous devons être capable de retrouver la probabilité du sous-ensemble de mondes possibles où un attribut prend une certaine valeur à partir de la distribution de probabilités donnée par la loi jointe *i.e.* l'ensemble des probabilités assignées à chaque monde possible. Cette opération s'appelle la marginalisation :

$P(X = i) = \Sigma_{A_1=\{v1_1,...v1_i\},...,A_m=\{vm_1,...,vm_j\}} P(X = i, A_1, ..., A_m)$

La valeur exacte d'une probabilité conditionnelle quelconque d'un réseau bayésien se calcule comme suit[22] :

Soit G=(X,E) un réseau bayésien défini sur $X = \{X_1, ..., X_n\}$ avec $n \in N$, E un ensemble d'arcs et $P(X_1 = v_1|X_2 = v_2, ..., X_j = v_j)$ où $j \leq n$ et $v_1, v_2, ..., v_j \in \{0, 1\}$, la probabilité recherchée :

1. $P(X_1 = v_1|X_2 = v_2, ..., X_j = v_j) = \frac{P(X_1=v_1,X_2=v_2,...,X_j=v_j)}{P(X_2=v_2,...,X_j=v_j)}$ par définition des probabilités conditionnelles
2. $\frac{P(X_1=v_1,X_2=v_2,...,X_j=v_j)}{P(X_2=v_2,...,X_j=v_j)} = \frac{\Sigma_{a_{j+1},...,a_n\in\{0,1\}} P(X_1=v_1,X_2=v_2,...,X_j=v_j,X_{j+1}=a_{j+1},...,X_n=a_n)}{\Sigma_{a_1,a_{j+1},...,a_n\in\{0,1\}} P(X_2=v_2,...,X_j=v_j,X_1=a_1,X_{j+1}=a_{j+1},...,X_n=a_n)}$ par marginalisation

valeur y et ...

[22] On suppose que les variables sont toutes binaires pour simplifier la présentation du calcul. Il peut être facilement généralisé à des variables d'arité quelconque.

3. $\frac{\Sigma_{a_{j+1},...,a_n \in \{0,1\}} P^{Simplifiee}(X_1=v_1,X_2=v_2,...,X_j=v_j,X_{j+1}=a_{j+1},...,X_n=a_n)}{\Sigma_{a_1,a_{j+1},...,a_n \in \{0,1\}} P^{Simplifiee} P(X_2=v_2,...,X_j=v_j,X_1=a_1,X_{j+1}=a_{j+1},...,X_n=a_n)}$
   où $P^{Simplifiee}(X_1 = v_1, X_2 = v_2, ..., X_j = v_j, X_{j+1} = a_{j+1}, ..., X_n = a_n)$
   et $P^{Simplifiee} P(X_2 = v_2, ..., X_j = v_j, X_1 = a_1, X_{j+1} = a_{j+1}, ..., X_n = a_n)$ sont les probabilités jointes simplifiées grâce aux indépendances décrites graphiquement par le réseau

Les probabilités de l'équation 3 peuvent être calculées grâce aux paramêtres du réseau et donnent une solution à la probabilité conditionnelle recherchée.

Inférence et révision des conclusions sur un exemple

Ce matin notre chef de projet se demande si Léo et Jean sont présents pour se joindre à lui pour boire un café. Sans plus d'informations que ses connaissances *a priori*, notre chef de projet ne peut que conclure la valeur *a priori* pour Jean et Léo d'être en retard, *i.e.* la croyance habituelle qu'il a, chaque matin en arrivant, de trouver Léo et Jean dans leur bureau. Pour déterminer la probabilité de trouver Jean dans son bureau à l'heure d'ouverture, compte tenu de nos *a priori* sur ses problèmes de réveil et de RER, nous devons calculer $P(Retard\_Jean = Non)$.

En marginalisant $P(Retard\_Jean = Non)$ et en utilisant la loi jointe simplifiée, nous obtenons :

$P(Retard\_Jean = Non) = \Sigma_{i,j,k=\{Oui,Non\}} P(Reveil\_Probleme = i) P(RER\_Probleme = j) P(Retard\_Jean = Non | Reveil\_Probleme = i, RER\_Probleme = j) P(Retard\_Leo = k | RER\_Probleme = j)$

Nous connaissons toutes les probabilités nécessaires au calcul de $P(Retard\_Jean = Non)$ grâce aux paramètres du RB, soit

$P(Retard\_Jean = Non) = (0.4 + 0.2 + 0 + 0.9) * (0.4 + 0.2 + 0 + 0.1) *$
$(0.4 + 0.8 + 0.2 + 0.01) * (0.4 + 0.8 + 0.2 + 0.99) * (0.6 + 0.2 + 0.1 + 0.9) *$
$(0.6 + 0.2 + 0.1 + 0.1) * (0.6 + 0.8 + 0.9 + 0.01) * (0.6 + 0.8 + 0.9 + 0.99) = 0.51376$

Le matin suivant notre chef de projet a rendez-vous en fin de matinée sur Paris et souhaite emprunter le RER pour s'y rendre mais il ignore s'il y a des perturbations sur la ligne. En arrivant notre chef de projet passe devant le bureau de Jean. Ce dernier étant vide, notre chef de projet en conclut que Jean est encore une fois en retard. Cette nouvelle information incite notre chef de projet à augmenter les probabibilités d'un disfonctionnement sur le RER ou d'un problème du réveil de Jean. Pour déterminer avec plus de certitude s'il s'agit d'un disfonctionnement notre chef de projet décide de vérifier la présence de Léo. Si ce dernier n'est pas dans son bureau alors c'est certainement un problème de RER. Dans le cas contraire, Jean a encore eu un problème de réveil !

Maintenant essayons de raisonner avec notre RB comme notre chef de projet l'a fait et vérifions que les conclusions du RB sont conformes à notre intuition. Dans notre exemple le chef de projet se demande si le RER subit encore des perturbations. En temps normal il estimerait cette possibilité à 20%. Mais, en arrivant, il a observé le retard de Jean. Il peut alors exploiter cette nouvelle information pour réviser son estimation du retard du RER. Notre chef de projet recherche la probabilité

$P(RER\_Probleme = Oui | Retard\_Jean = Oui)$

Par définition des probabilités conditionnelles nous savons que :

$P(RER\_Probleme = Oui|Retard\_Jean = Oui) = \frac{P(RER\_Probleme=Oui,Retard\_Jean=Oui)}{P(Retard\_Jean=Oui)}$

La probabilité *P(Retard_Jean=Oui)=1-P(Retard_Jean=Non)* nous est donnée par le calcul précédent. Nous devons marginaliser *P(RER_Probleme=Oui,Retard_Jean=Oui)* pour pouvoir calculer la probabilité recherchée,

$P(RER\_Probleme = Oui, Retard\_Jean = Oui) =$
$\Sigma_{i,j=\{Oui,Non\}} P(Reveil\_Probleme = i)P(RER\_Probleme = Oui)P(Retard\_Jean =$
$Oui|Reveil\_Probleme = i, RER\_Probleme = Oui)P(Retard\_Leo = j|$
$RER\_Probleme = Oui) = (0.4 + 0.2 + 1 + 0.9) * (0.4 + 0.2 + 1 + 0.1)*$
$(0.6 + 0.2 + 0.9 + 0.9) * (0.6 + 0.2 + 0.9 + 0.1) = 0.188,$

d'où *P(RER_Probleme=Oui|Retard_Jean=Oui)=0.3866*. Un calcul identique nous montrerait que *P(Reveil_Probleme=Oui|Retard_Jean=Oui)=0.68*, sur la seule évidence que Jean est en retard et de nos *a priori*, on en conclut qu'il a plus certainement eu une nouvelle fois un problème de réveil. Supposons maintenant que notre chef de projet ait aussi observé l'absence de Léo, la probabilité qu'il souhaite évaluer est

P(RER_Probleme=Oui|Retard_Jean=Oui,Retard_Leo=Oui)=0.98

et il en conclut qu'il y a cette fois-ci un problème de RER (toutefois Jean peut avoir eu aussi un problème de réveil *P(Reveil_Probleme=Oui|Retard_Jean=Oui,Retard_Leo=Oui)=0.43*).

L'inférence avec des observations incertaines

Il arrive parfois que nos observations soient, elles aussi, incertaines. La lecture d'une gradation sur un thermomètre ménager est souvent imprécise, nous pouvons donc affirmer que la température de la pièce est de 20°C avec une certitude de 0.97, ou encore être certain qu'un attribut ne prend pas la valeur v_1 et v_2 mais ignorer s'il a la valeur v_3 ou v_4. Ce type d'évidence que l'on nomme "observation vraisemblable" ou *soft evidence* s'oppose à la *hard evidence*. Pour insérer ces observations partielles ou négatives dans le réseau, une méthode simple consiste à ajouter une VA mesurant la fiabilité de nos observations pour l'attribut qu'elle influence. Nous la distinguons par une couleur différente dans la figure 4.2. Cette VA est d'une nature identique à celles des autres VA du réseau et elle est utilisée de la même manière dans les calculs d'inférences. Mais sa "valeur sémantique" est différente : elle n'exprime pas la probabilité pour un atribut de prendre une certaine valeur, elle exprime notre confiance dans la probabilité que cet attribut prenne cette valeur. [Fabiani, 1996], qui discute en détail ce point théorique, emploie l'expression *probabilité de second ordre* pour désigner les *soft evidence*.

Pour prolonger notre exemple précédent, lorsque Paul arrive au bureau il est mal réveillé. En passant devant le bureau de Jean, il a cru l'apercevoir mais peut très bien l'avoir confondu avec un de ses collègues. Il reste toutefois deux fois plus sûr de l'avoir aperçu que de l'avoir confondu avec quelqu'un d'autre. Le réseau de la figure 4.2 a été modifié pour insérer ces observations partielles. Une inférence identique aux inférences précédentes permet de calculer la nouvelle probabilité *P(Retard_Jean=Oui|Fiabilite_Observation=Vraisemblable)=0.66* et d'inférer que *P(Reveil_Probleme=Oui)=0.5*.

Les inférences que nous avons réalisées dans les exemples précédents nous montrent que lorsqu'une croyance sur la valeur d'une VA vient à être modifiée, cela entraîne une modification de nos croyances sur les valeurs de certaines autres variables. Le fait de savoir que Jean est en retard augmente les probabilités d'un problème de réveil et de RER. Le sens de la propagation de l'information dans le réseau peut être prévue à partir d'une contrainte purement

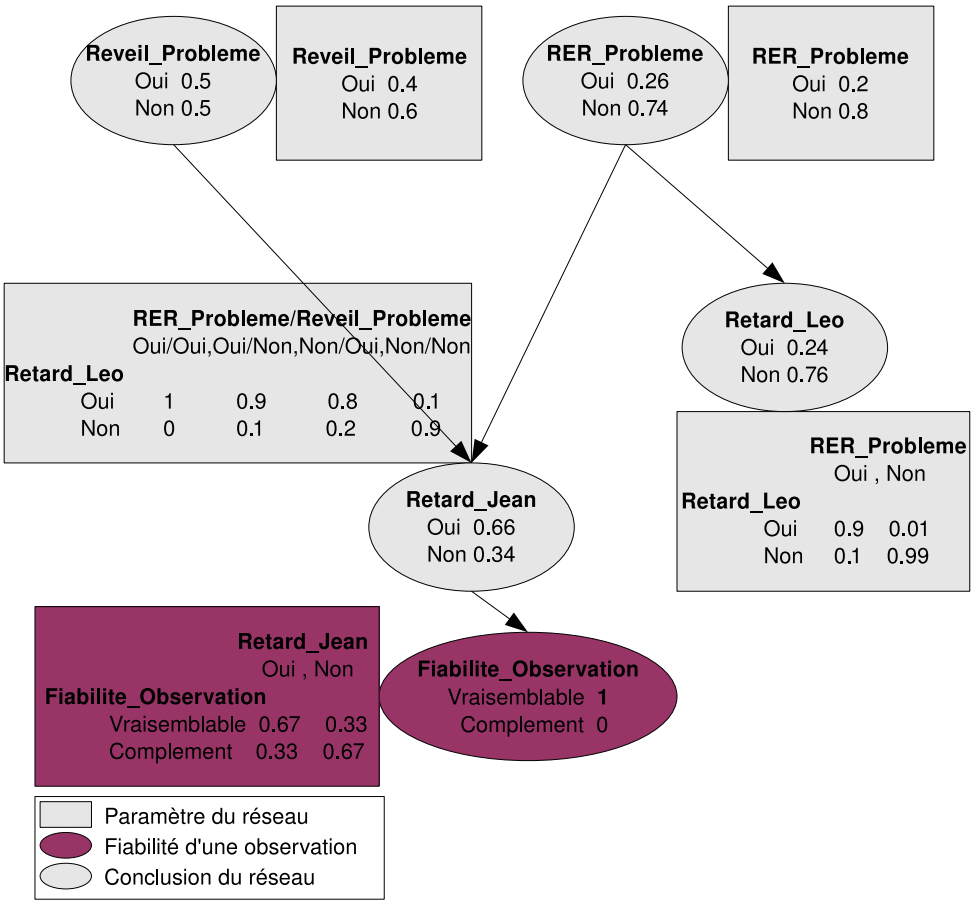

FIG. 4.2 – Exemple d'une observation vraisemblable

graphique. Cette contrainte, que l'on nomme *d-séparation*, en distinguant les VA dont les probabilités sont révisées et celles qui restent inchangées lorsqu'une nouvelle information vient à être connue, facilite les calculs lors de l'inférence. Toutefois, il suffit que le nombre de VA et leur connexions augmentent pour que les calculs nécessaires à l'inférence deviennent difficiles à réaliser en un temps raisonnable. Plusieurs algorithmes d'inférence, dont la présentation est hors de notre propos ici, coexistent. Une présentation des algorithmes classiques peut être trouvée dans [Jensen & Nielsen, 2007].

4.1.3 L'apprentissage d'un réseau bayésien

Pour définir entièrement un réseau bayésien nous avons besoin de connaître les VA qui le composent, la structure du graphe et enfin les valeurs des paramètres. L'apprentissage de chacun de ses composants peut être réalisé automatiquement avec un succès qui dépend de la quantité et de la qualité des données d'apprentissage.

Apprentissage des paramètres

Les paramètres sont certainement les composants les plus simples à apprendre. Plaçons-nous dans le cas où la structure et les variables sont toutes connues. Les probabilités conditionnelles peuvent-être approximées soit par expertise, soit par analyse d'un ensemble de données.

Lorsqu'un expert est disponible, il peut être interrogé pour fixer les valeurs des probabilités. Des outils peuvent être employés pour guider son évaluation et atténuer les biais que l'expert pourrait introduire. Un outil connu et facile à utiliser selon [Naïm *et al.*, 2004] est l'échelle de probabilité reproduite dans la figure 4.3.

L'expertise est limitée par la complexité du réseau. A titre d'exemple, pour une probabilité conditionnelle $P(Y|X_1, ..., X_n)$ où $Y, X_1, ..., X_n$ sont des variables binaires (*i.e.* qui n'admettent que deux valeurs), il faut déterminer 2^n valeurs. Une solution possible pour réduire le nombre de valeurs qu'il faut demander à l'expert consiste à calculer une probabilité conditionnelle complexe $P(Y|X_1, ..., X_n)$ à partir de probabilités conditionnelles plus simples :

P(Y=non$|X_1$*=oui),...,P(Y=non*$|X_n$*=oui)*[23].

Cette méthode porte le nom de OU-bruité, elle a été initialement conçue pour des variables binaires et étendue pour des variables multivaluées (on parle alors de *generalized noisy-OR gate*). L'expertise peut être séparée entre plusieurs experts mais cette solution pose un autre problème : les probabilités des experts doivent être cohérentes entre elles[24]. Différentes méthodes existent pour unifier les expertises en tenant compte de la fiabilité respective des experts [Paass, 1988],[Naïm *et al.*, 2004].

Les données disponibles sont une autre source utilisable pour établir les paramètres d'un RB. Nous entendons par "données" une table de base de données où un enregistrement assigne

[23] On suppose que les X_i sont indépendants entre eux.

[24] [Paass, 1988] montre la contradiction cachée derrière des probabilités données par expertises
- $P(B = oui|A = oui) = 0.1$,
- $P(A = oui) = 0.9$,
- $P(B = oui) = 0.9$,
- $P(B = non) \geq P(B = non \cap A = oui) = P(B = non|A = oui)P(A = oui) = 0.81$

or par hypothèse $P(B = non) = 0.1$.

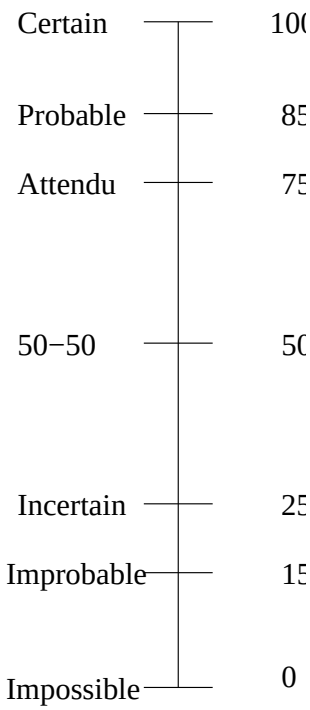

FIG. 4.3 – Une aide pour l'expert : l'echelle de probabilité

une valeur pour chaque VA du réseau. Bien sûr, certaines valeurs peuvent être erronées ou manquantes et certaines autres, possibles mais rares, sont absentes de la base lorsque cette dernière contient trop peu de données. Ces critères influencent le choix d'une méthode d'apprentissage particulière. Lorsque les données sont complètes la méthode la plus simple consiste à estimer les probabilités à partir des fréquences observées sur les données, cette approche est appelée le *maximum de vraisemblance* :

$P(X = i|A_1 = j_1, ..., A_n = j_n) = \frac{\sharp(X=i,A_1=j_1,...,A_n=j_n)}{\sharp(A_1=j_1,...,A_n=j_n)}$, où $Parent(X) = A_1, ..., A_n$ et $\sharp(X = i, A_1 = j_1, ..., A_n = j_n), \sharp(A_1 = j_1, ..., A_n = j_n)$ les nombres d'enregistrements de la base de données où les variables $X, A_1, ..., A_n$ ont la valeur $i, j_1, ..., j_n$ respectivement.

Si notre chef de projet n'avait fait confiance qu'à sa propre expertise pour déterminer les probabilités précédentes, il aurait pu renseigner les données de la table 4.1 en relevant chaque matin, sur une période d'une semaine, les retards de Jean et de Léo ainsi que leurs motifs réciproques. Les probabilités conditionnelles se calculent alors comme suit :

$P(Retard\_Jean = Non|Reveil\_Probleme = Non, RER\_Probleme = Non) = \frac{4}{5}$,
$P(Retard\_Jean = Non|Reveil\_Probleme = Oui, RER\_Probleme = Non) = \frac{1}{1}$,...

Lorsque la quantité de données disponibles est trop petite, certains cas de figures peuvent être insuffisament représentés, voire être absents. Dans ce cas, les paramètres calculés avec l'approche du maximum de vraisemblance sont non représentatifs de la population. Le principe de l'estimation bayésienne évite ce biais. Ce principe est mis en oeuvre dans les méthodes du maximum *a posteriori* et l'espérence *a posteriori*. Dans cette approche on suppose un paramétrage *a priori* que l'on révise selon les données observées. On peut interpréter cette approche

Reveil_Probleme	RER_Probleme	Retard_Jean	Retard_Leo
Non	Non	Non	Non
Non	Non	Oui	Non
Oui	Non	Non	Non
Non	Non	Non	Non
Non	Non	Non	Non
Non	Non	Non	Non
Oui	Oui	Oui	Oui

TAB. 4.1 – Exemple de données pour l'apprentissage des paramètres

comme un ajout de données virtuelles, estimées par un expert, à l'ensemble de données réelles [Jensen & Nielsen, 2007].

Maximum a posteriori :
$P(X = i|A_1 = j_1, ..., A_n = j_n) = \frac{\sharp(X=i,A_1=j_1,...,A_n=j_n)+\sharp'(X=i,A_1=j_1,...,A_n=j_n)-1}{\sharp(A_1=j_1,...,A_n=j_n)+\sharp'(A_1=j_1,...,A_n=j_n)-1}$
Espérance a posteriori :
$P(X = i|A_1 = j_1, ..., A_n = j_n) = \frac{\sharp(X=i,A_1=j_1,...,A_n=j_n)+\sharp'(X=i,A_1=j_1,...,A_n=j_n)}{\sharp(A_1=j_1,...,A_n=j_n)+\sharp'(A_1=j_1,...,A_n=j_n)}$,
où $Parent(X) = A_1, ..., A_n$ et $\sharp(X = i, A_1 = j_1, ..., A_n = j_n), \sharp(A_1 = j_1, ..., A_n = j_n)$ sont les données réelles et $\sharp'(X = i, A_1 = j_1, ..., A_n = j_n), \sharp'(A_1 = j_1, ..., A_n = j_n)$ des données correctives ajoutées par l'expert

Au regard des données de notre exemple jouet, Léo n'arrive jamais à l'heure lorsqu'il y a un problème de RER *P(Retard_Leo=Non|RER_Probleme=Oui)=*$\frac{0}{1}$. Notre chef de projet qui a déjà vu le cas inverse se produire, décide de modifier le calcul de ses paramètres en ajoutant les cas :

$\sharp$*(Retard_Leo=Non|RER_Probleme=Oui)=1* et $\sharp$*(Retard_Leo=Oui|RER_Probleme=Oui)=9*, soit l'espérance *a posteriori* de notre paramètre :

P(Retard_Leo=Non|RER_Probleme=Oui)=(0+1)/(1+10)=0.09

Quand la valeur d'une ou de plusieurs VA est inconnue pour certaines occurrences des données, il existe des algorithmes permettant de compléter les valeurs manquantes dont l'un des plus connus est l'algorithme *Expectation Maximisation* (EM). Nous ne détaillerons pas ces algorithmes car, bien que nos données soient incomplètes, nous avons choisi une solution plus simple (voir la section 6.1.2) pour l'apprentissage des paramètres.

Apprentissage du réseau

L'apprentissage du réseau est une étape plus complexe que l'apprentissage des paramètres. Deux difficultés se posent. Avant de rechercher la structure il faut découvrir les variables latentes. Nous pouvons dans certains cas identifier des corrélations entre certains faits du monde et ignorer la cause de cette influence. Une corrélation existe entre le fait que les parisiens s'abritent sous leur parapluie et que la seine déborde, mais aucun des faits n'est cause de l'autre. Pour comprendre le phénomène il faut rajouter la variable cachée *Pluie*.

Une fois toutes les variables pertinentes déterminées, la recherche de la structure n'est pas plus simple. Une analyse exhaustive de toutes les structures possibles entre *n* nœuds est impossible en raison du nombre de ces structures. On peut réduire l'espace de recherche de l'analyse en introduisant la notion de représentant d'une classe d'équivalence de Markov. On interprète

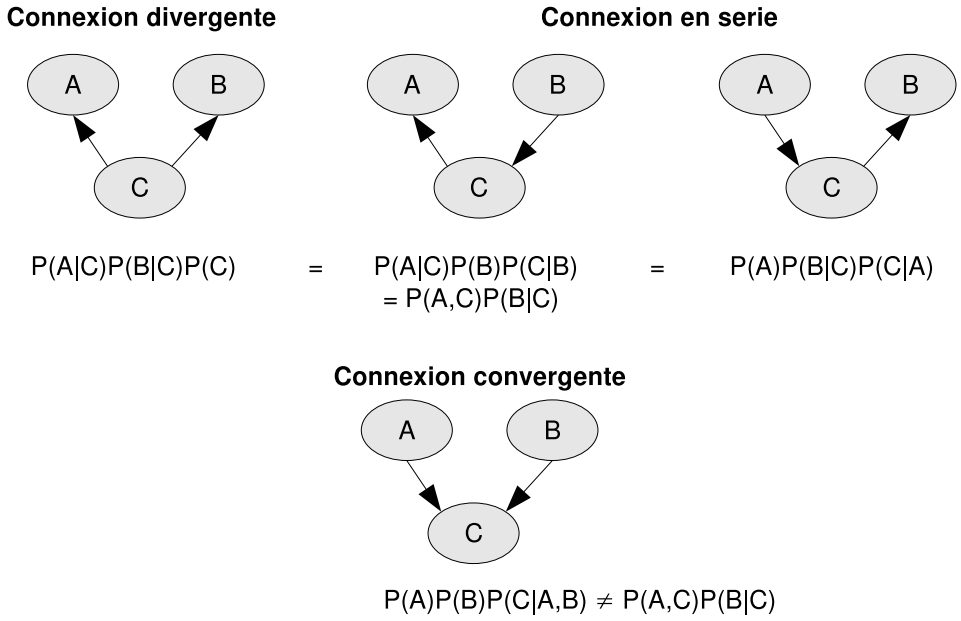

FIG. 4.4 – Equivalences de Markov

habituellement l'arc reliant deux nœuds d'un réseau bayésien comme un lien de causalité entre deux VA. Cette interprétation a pour but de faciliter le travail de modélisation de la structure réalisé par l'expert. Mais les liens d'un réseau sont rarement des liens causaux. La causalité est une notion difficile à définir. Aucune définition n'est partagée unaniment [Kayser & Levy, 2004]. La définition de l'implication logique est la plus simple : A cause B lorsque l'ensemble des mondes où B est vrai est inclus dans celui où A l'est. Or la probabilité conditionnelle est une mesure de la corrélation entre deux faits *i.e.* du nombre de mondes où A et B sont vrais simultanément,

$P(B|A)P(A) = P(A \cap B) = P(A|B)P(B)$[25].

En utilisant l'égalité des probabilités conditionnelles, on démontre l'équivalence des probabilités jointes des trois premières structures de la figure 4.4 mais pas de la dernière. Lorsqu'un expert modélise une structure précise sur la connaissance des relations causales de son domaine, il modélise en réalité un représentant de la classe de Markov. A la condition que l'intégrité des connexions convergentes (voir la quatrième structure de la figure 4.4) soit respectée et qu'aucune autre connexion convergente ne soit introduite, il est possible de travailler sur des structures différentes mais équivalentes à celle de l'expert. Car ces structures modélisent une loi jointe identique.

Les algorithmes pour l'apprentissage de la structure se divisent en deux catégories. Les premiers recherchent les indépendances conditionnelles à l'aide de tests statistiques, les seconds assignent des scores aux structures. Des versions modifiées incluent une étape de découverte des variables latentes. Nous ne détaillerons pas davantage la présentation de ces algorithmes n'ayant

[25] La probabilité conditionnelle n'exprime la causalité au sens de l'implication logique que dans le cas particulier où A est vrai dans tous les mondes possibles

$P(A \rightarrow B) = P(B|A)$

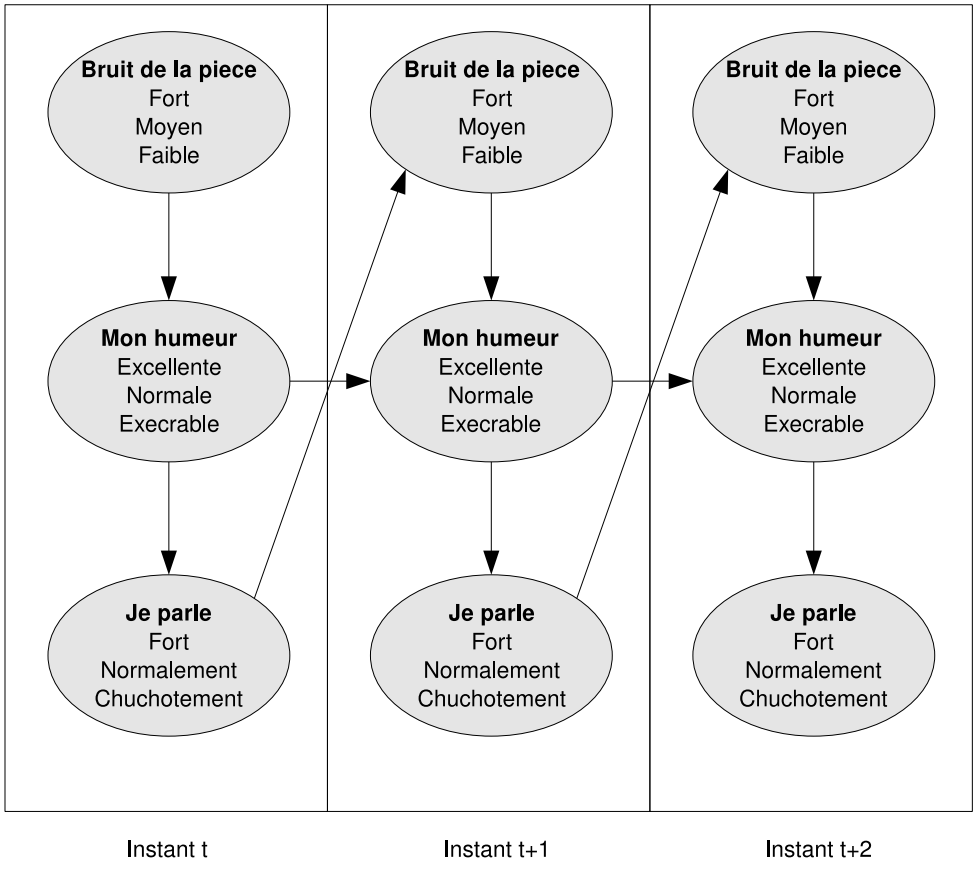

FIG. 4.5 – Exemple d'un réseau bayésien dynamique

pu, faute de temps, les appliquer à nos problèmes.

4.2 Les limites du modèle des réseaux bayésiens

4.2.1 Les réseaux bayésiens dynamiques

Le formalisme des réseaux bayésiens est un modèle expressif mais il souffre d'une contrainte handicapante : la structure du réseau doit être connue et rester figée durant toute l'inférence. Cette contrainte, nous le verrons dans le chapitre 6, nous interdira une modélisation correcte du problème du choix de l'antécédent. Pour lever cette contrainte nous serons obligé d'employer une extension du formalisme que nous présentons ici. Le modèle des réseaux bayésiens dynamiques (RBD) a été initialement conçu pour intégrer la notion de temps. Un réseau bayésien modélise l'état du monde à un instant *t* donné. Une occurrence du réseau est reproduite pour chaque instant suivant l'instant *t* avec l'ajout des liens d'influence entretenus par les variables d'une occurrence d'un instant avec les variables de l'occurrence de l'instant suivant (ou des instants suivants). Le réseau de la figure 4.5 modélise naïvement l'évolution du fond sonore d'une pièce sur trois instants.

Lorsque je rentre dans une pièce, le fond sonore de celle-ci va modifier mon humeur et, de mon humeur dépend la façon avec laquelle je vais parler dans la pièce, par exemple si le fond sonore est élevé cela m'incitera à élever la voix. Les influences des trois VA restent identiques à l'instant qui suit mon entrée mais deux d'entre elles sont influencées par les VA de l'instant de mon entrée dans la pièce :

- mon humeur actuelle dépend de mon humeur précédente car je ne passe pas d'une humeur exécrable à une humeur joyeuse sans transition,
- le volume sonore de ma voix vient modifier (ou non) le contexte sonore de la pièce.

4.2.2 Les réseaux d'inférence

La contrainte principale des réseaux bayésiens "statiques" ou dynamiques est que ceux-ci ne doivent pas contenir de circuit. Pour certains problèmes, comme ceux des diagnostics médicaux, il serait utile d'exprimer deux connaissances incertaines distinctes telles que

la présence du symptôme A trahit la maladie B avec une probabilité π_1 et

la maladie B implique la présence du symptôme A avec une probabilité π_2.

Ce qui n'est pas possible avec un réseau bayésien. Cette limitation est dûe à la sémantique causale associée aux arcs du réseau. Lorsqu'un fait A est cause de B, en général B n'est pas cause de A. L'introduction d'un circuit dans le réseau produit une boucle de renforcement qui n'est justifiée par aucune observation :

- *P(maladie|symptome)=0.06* et *P(maladie|absence symptome)=0.04*,
- *P(symptome|maladie)=0.12* et *P(symptome|absence maladie)=0.28*,

Si nous partons avec l'*a priori P(maladie)=0.3* et *P(symptome)=0.6*, on calcule successivement lors de l'étape d'inférence :

P(maladie)=0.052, P(symptome)=0.556 puis

P(maladie)=0.0511, P(symptome)=0.271 puis...

Cette contrainte est levée dans le modèle des réseaux d'inférence. Différents travaux proposent des méthodes de calculs évitant le renforcement circulaire parmi lesquels [Paass, 1988] ou les travaux cités par [Fabiani, 1996]. Mais le fait de relâcher cette contrainte a un coût, les calculs nécessaires pour interroger le réseau sont plus complexes, et il n'est pas certain que cette contrainte doive être relâchée pour modéliser correctement un problème d'annotation. Nous avons, en conséquence, fait le choix de conserver cette contrainte pour satisfaire notre premier objectif qui est de déterminer l'apport du traitement de l'incertitude pour un système d'annotations.

4.3 Des cas particuliers célèbres

4.3.1 Le classifieur bayésien naïf

Le classifieur bayésien naïf (BN) est toujours un modèle de référence en raison de sa simplicité et de ses bonnes performances [Domingos & Pazzani, 1996]. Le BN est un réseau bayésien particulier. Contrairement à un RB où l'on s'intéresse aux valeurs de plusieurs VA impliquées dans un réseau d'influence, dans un BN, on s'intérresse à la valeur d'une unique variable influencée par un ensemble de VA observables dont on ignore volontairement les relations d'in-

fluence. Techniquement, un BN est un RB où la structure est un arbre de profondeur 1 avec ses feuilles mutuellement indépendantes. Le BN est employé pour les tâches de classification. Les valeurs de la racine représentent les classes pouvant subsumer une donnée quelconque, les feuilles, les attributs décrivant la donnée. Les paramètres du BN sont estimés à partir d'un ensemble de données étiquetées. Lorsqu'une donnée inconnue est présentée au BN, les valeurs des attributs décrivant cette donnée sont posées comme observations dans le réseau et, au terme de l'inférence, la classe qui obtient la plus grande probabilité est affectée à la donnée. Nous commentons l'exemple d'un BN filtrant les *spam* dans la section 4.5.

4.3.2 Les modèles de Markov cachés

Les réseaux bayésiens dynamiques ont été largement appliqués pour la reconnaissance de la parole sous la forme simplifiée des modèles de Markov Cachés (MMC) (l'article [Rabiner, 1989] est une référence de ce domaine). Les MMC permettent de modéliser une séquence aléatoire d'événements observables régit par un ensemble d'états cachés. L'interrogation d'un MMC est multiple. On peut vouloir estimer la probabilité d'émission d'une séquence d'évènements observés, la probabilité de l'évènement à suivre connaissant les évènements passés ou encore d'induire la suite la plus probable des états cachés parcourus lors de l'émission d'une séquence. Un MMC est classiquement défini par un triplet (A,B,π) où

- A est une matrice de probabilités de transition entre les états cachés S=$\{s_1, ..., s_n\}$ qui régissent le processus
- B est la matrice de probabilités des observations
- π est un vecteur où chaque scalaire est la probabilité pour que le processus débute par l'état associé au scalaire

Il peut être reconstruit comme un RBD composé d'un ensemble de VA *cachées* et *observables*. La figure 4.6, copiée sur [Denoyer, 2004], montre un MMC modélisant le processus météorologique. Le temps que l'on peut observer *Pluie, Soleil* et *Orage* est conditionné par les 4 états cachés du climat *Temps anticyclonique, Ciel de traîne, Dépression* et *Orageux*. Le climat pouvant passer d'un état à un autre selon certaines probabilités. La figure 4.7 complète la figure du précédent MMC en représentant le RBD qui lui est associé.

4.4 Une famille concurrente : les modèles discriminants

4.4.1 Définition des modèles discriminants

Notre présentation des modèles génératifs serait imcomplète si elle ne mentionnait pas, même brièvement, la famille des modèles concurrents des modèles génératifs : les modèles discriminants. Un modèle discriminant n'est pas à proprement parler un modèle de raisonnement sur l'incertain. Il s'agit plutôt d'un classifieur dont les bonnes performances et la facilité d'utilisation en font un modèle très largement employé en apprentissage. Le principe général des méthodes discriminatives consiste à projeter les données dans un espace géométrique dans le but de trouver un hyperplan séparant ces données selon leur classe d'appartenance. La figure 4.8 montre un espace géométrique de dimension 2 où les exemples ont pu être séparés linéairement en deux classes par la droite d'équation $a_0 + a_1x_1 + a_2x_2 = 0$.

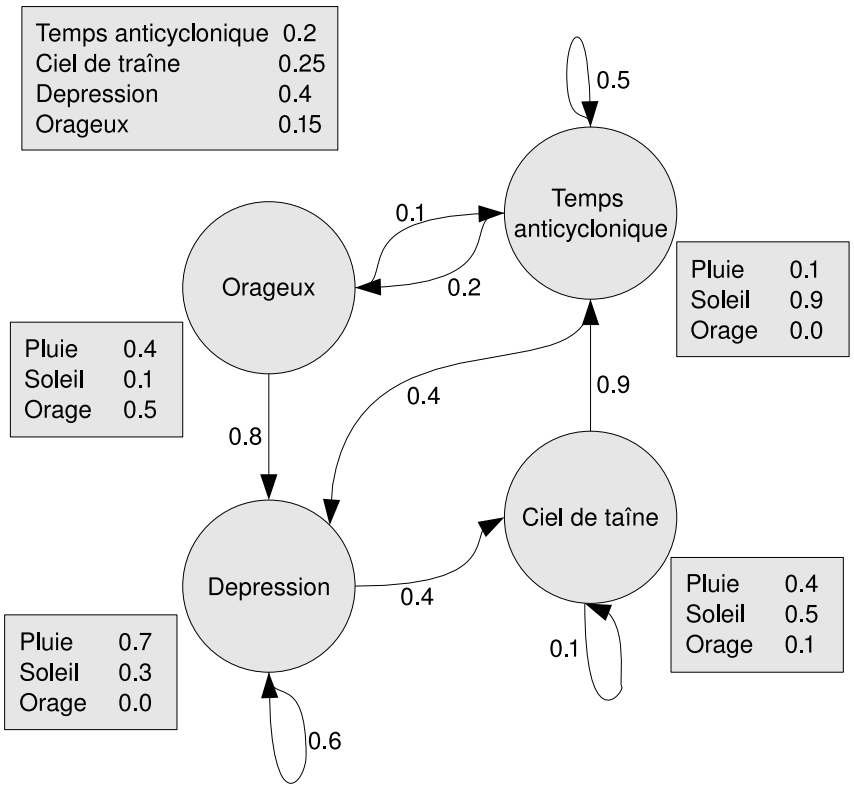

FIG. 4.6 – Un modèle Markov caché modélisant le processus météorologique

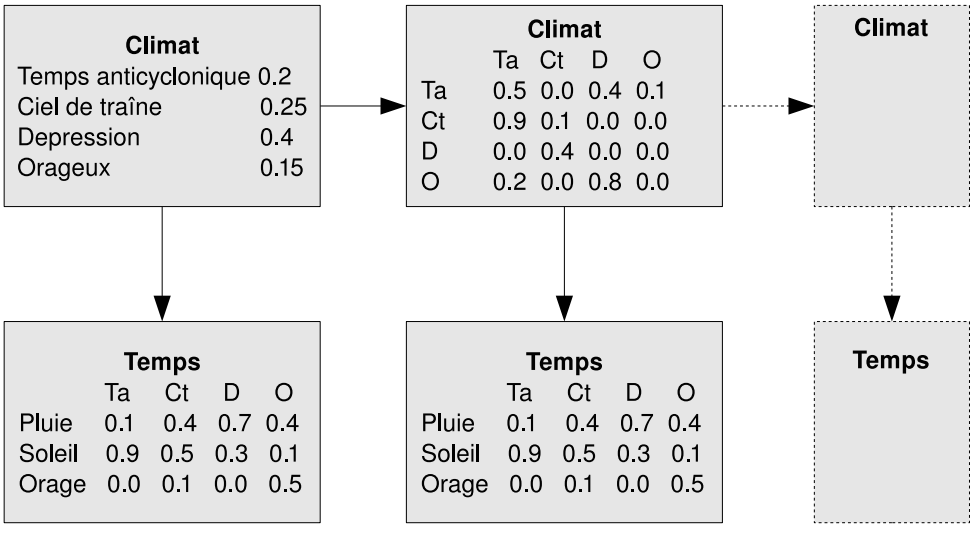

FIG. 4.7 – Un réseau bayésien dynamique modélisant le processus météorologique

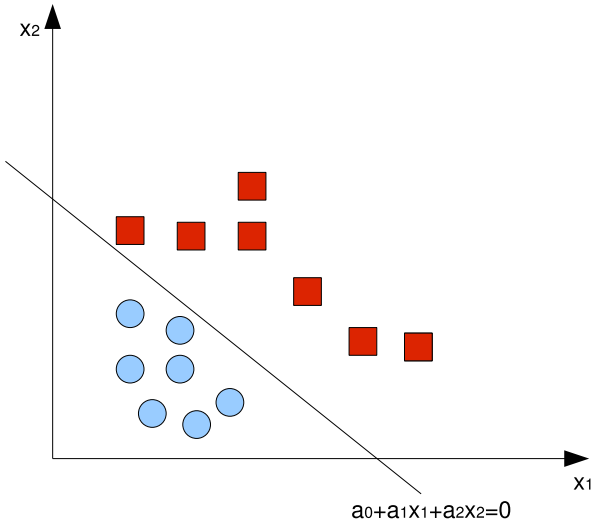

FIG. 4.8 – Exemple de données linéairement séparables

Des modèles ont été proposés pour des problèmes plus complexes, comme, par exemple, des problèmes où les données doivent être séparées en plus de deux classes, ou encore des problèmes où la séparation linéaire est impossible sans changer la dimension de l'espace géométrique. Le modèle des séparateurs à vaste marge (ou *Support Vector Machine (SVM)*) reste le modèle discriminant qui remporte, à l'heure actuelle, le plus grand succès et continue d'être le plus étudié [Cornuéjols & Miclet, 2002].

4.4.2 Modèles discriminants *vs* modèles génératifs

Même si les méthodes discriminantes affichent, en général, de meilleures performances que les méthodes génératives, la performance n'est pas l'unique critère dont il faut tenir compte dans le choix d'un modèle. Les auteurs de [Naïm *et al.*, 2004] avancent plusieurs arguments en faveur des réseaux bayésiens. Nous résumons ici les arguments principaux.

Pour de nombreux domaines d'application, les connaissances nécessaires à la conception du modèle souhaité sont souvent insuffisantes et hétérogènes. Un des avantages des réseaux bayésiens qui est mis en avant est une acquisition et une modélisation facile de ces connaissances au sein du modèle. Leur expression s'effectue intuitivement par de simples liens de causalité et le pouvoir d'expression très abstrait des variables aléatoires permet de les intégrer au sein d'un modèle unique quelle que soit la nature de ces connaissances : règles logiques, données numériques d'apprentissage, données statistiques ou encore probabilités subjectives d'experts.

La représentation graphique des connaissances et de leurs interactions peut être comprise rapidement par un utilisateur, contrairement aux modèles discriminants dont le fonctionnement reste opaque pour les non initiés. De plus, le réseau obtenu se prête à l'interrogation d'un utilisateur soucieux de diagnostiquer les raisons d'une mauvaise décision prise par le système ou encore de prévoir son comportement dans des situations rares ou absentes dans les données d'études disponibles.

Enfin, le choix du réseau bayésien aux dépens d'un modèle discriminant n'est pas irréversible. En effet, il est toujours possible de "traduire" un réseau bayésien en un SVM et de bénéficier des performances de ce dernier modèle. Dans sa thèse, [Denoyer, 2004] détaille les différentes procédures possibles et résume dans [Denoyer & Gallinari, 2004] les gains obtenus sur une tâche de recherche d'information.

4.5 L'utilisation des réseaux bayésiens dans le TAL

Le plaidoyer de la section précédente a été rédigé par des chercheurs travaillant dans le domaine de l'apprentissage automatique [Naïm *et al.*, 2004]. La défense du modèle repose donc logiquement sur des arguments généraux de ce domaine. Dans cette section, au travers de divers travaux qui ont employé les réseaux bayésiens, ou des extensions du modèle, pour des problèmes de TAL, nous soulignons les caractéristiques qui en font un très bon modèle d'inférence pour le domaine du TAL.

La communauté s'entend pour dater les premiers travaux sur les réseaux bayésiens dans les années 1980 et les premières applications dix ans plus tard. C'est donc un modèle récent, si l'on fait exception du classifieur bayésien naïf et des MMC, qui commence seulement à être employé par la communauté du TAL, malgré les intérêts qu'il présente. Ses propriétés lui permettent d'exploiter les annotations d'entrée en dépit des leurs imperfections.

4.5.1 Une adaptation automatique au domaine du corpus

Le premier article que nous présentons, [Sahami *et al.*, 1998], décrit un système de Recherche d'Information (RI) spécialisé pour la distinction des *spams* et des *e-mails*. Même si les auteurs n'emploient pas, à proprement parler, un réseau bayésien mais un classifieur bayésien naïf, ils soulignent clairement le premier intérêt d'un classifieur bayésien. L'adaptation automatique du classifieur au domaine du corpus réduit l'impact de l'erreur d'estimation dans la décision.

Les auteurs partent du constat que les premiers filtres *anti-spam* qui reposaient sur un ensemble de règles logiques rigides étaient mal adaptés pour cette tâche obligeant l'utilisateur à corriger les règles, une tâche longue et fastidieuse. Non seulement les *spammeurs* contournaient facilement les règles standard mais ces dernières ne pouvaient satisfaire tous les utilisateurs. Une règle filtrant tout message comportant le mot "Viagra" convient certainement pour un étudiant en informatique mais pas pour un commercial de l'industrie pharmaceutique dont ce produit apparaît régulièrement dans la correspondance électronique. Le système doit être fondé sur un mécanisme capable d'adapter automatiquement la valeur des informations utilisées lors de la classification en fonction du corpus qui est construit selon les besoins de l'utilisateur. Plus précisément le mécanisme doit réviser ses connaissances au regard des caractéristiques d'un corpus d'un domaine particulier. La solution pour nos auteurs réside dans l'apprentissage des paramètres du classifieur bayésien naïf de la figure 4.9.

Leur système regroupe trois types d'attributs pour filtrer les *e-mails*. Le premier type d'attributs est un vocabulaire calculé automatiquement sur un corpus d'apprentissage et composé des 500 mots[26] qui apportent le plus d'information. Chaque VA Mot_i marque la présence ou

[26] Le nombre a été déterminé à partir des expériences et peut être modifié.

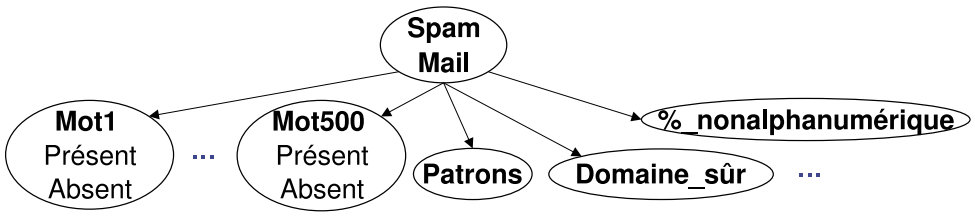

FIG. 4.9 – Un classifieur bayésien naïf pour le filtrage des *e-mails*

l'absence du i^{me} mot du vocabulaire dans un mail inconnu devant être classé. Le second type d'attributs est une courte grammaire qui reconnaît les expressions les plus employées par les *spammeurs* comme "FREE !". Enfin le dernier type d'attributs relève les caractéristiques générales du mail comme le pourcentage de caractères non-alphanumériques dans le sujet du mail ou encore le domaine de l'expéditeur et l'heure d'envoi de son mail. L'apprentissage des paramètres a été réalisé sur un corpus d'entrainement réel, 2593 *e-mails* et *spams* d'un utilisateur reçus en un an. Le corpus de test est, lui, composé des 222 mails hebdomadaires du même utilisateur. Un mail est rejeté comme *spam* si le classifieur chiffre une probabilité égale ou supérieure à 0.999 pour cette classe. Les auteurs préfèrent présenter un *spam* à l'utilisateur plutôt que de rejeter un *e-mail*. Les résultats obtenus sont satisfaisants car aucun *e-mail* n'est rejeté et 39 des 45 *spams* sont écartés par le système.

Nous revenons dans les expériences des chapitres suivants sur les raisons et le profit que l'on peut attendre de cette adaptation automatique pour limiter la chute des performances d'un système dû au changement de corpus.

4.5.2 La prise en compte de la fiabilité des informations

Les travaux de [Roth & Wen-tau, 2002] défendent un autre aspect des RB que nous avons déjà largement développé au début de ce chapitre : la possibilité de raisonner à partir de connaissances incertaines. On retrouve une intention similaire dans les travaux de [Ding & Peng, 2004] pour une tâche de modélisation et de fusion d'ontologies.

[Roth & Wen-tau, 2002] remarquent que les systèmes traditionnels d'extraction d'information traitent l'étiquetage des entités nommées séparément de celui des relations. Les erreurs de reconnaissance des entités se propagent donc à la reconnaissance des relations et, inversement, une erreur dans la reconnaissance d'une relation empêche de désambiguïser le type de ses arguments. Or, l'apparition des entités et celle de leurs relations ne sont pas indépendantes : dans des textes de faits divers par exemple, savoir que les entités x et y sont des noms de personnes renforce la probabilité qu'elles soient les arguments d'une relation *x est l'assassin de y*.

En reconnaissant les entités et leurs relations simultanément, [Roth & Wen-tau, 2002] montrent comment le RB peut tirer profit de cette dépendance mutuelle des attributs pour corriger les erreurs d'assignation de leurs valeurs.

Le système de [Roth & Wen-tau, 2002] a pour tâche d'extraire simultanément les arguments et les relations binaires *être_l'assassin_de* et *être_né_à* dans un petit corpus journalistique où sont mélangées 245 phrases contenant la relation *être_l'assassin_de*, 179 phrases avec la relation *être_né_à* et 502 phrases sans relation. Pour alléger les calculs du RB, ils simplifient le

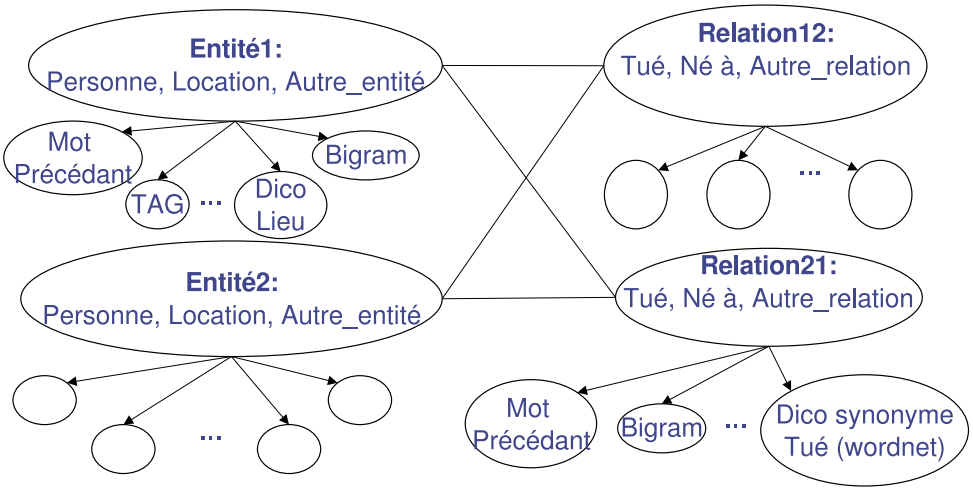

FIG. 4.10 – Un *loopy bipartite DAG* pour l'extraction simultanée des relations et de leurs arguments

problème en séparant le réseau. Chaque entité ou relation est reconnue indépendement par un classifieur. Les informations décrivant les EN et les relations sont classiques. Pour les EN, les auteurs mentionnent les mots précédant l'EN, leur étiquette morphosyntaxique et des dictionnaires de noms propres par exemple, et pour les relations les mots autour et entre les arguments ainsi que les synonymes de *Wordnet* des verbes *kill* et *born*. Les probabilités des classifieurs sont ensuite passées en paramètre au réseau de la figure 4.10 qui infère la relation la plus probable compte tenu des arguments.

Les auteurs veulent tenir compte des dépendances mutuelles entre les entités et leur relations. Ils considèrent une extension particulière des RB : les *loopy bipartite DAG* qui autorisent les circuits dans le réseau et pour lesquels un algorithme d'inférence approximatif a été proposé. La validation du modèle est obtenu en comparant les résultats des classifieurs indépendants et ceux du système complet. Les auteurs rapportent une petite baisse du rappel du système complet mais une augmentation de sa précision : les contraintes qu'il exprime corrigent les incohérences produites par les classifieurs indépendants.

4.5.3 Un biais de représentation réduit

La qualité la plus remarquée des RB par les chercheurs du TAL reste, sans doute, la facilité avec laquelle ils peuvent intégrer des informations hétérogènes. Une qualité précieuse lorsque l'on sait que le traitement d'un document nécessite un grand nombre d'annotations de complexités et de natures variées.

Dans sa thèse [Denoyer, 2004], l'auteur réfléchit sur la récente évolution de la recherche d'information. Les résultats principaux que nous discutons ici se trouvent résumés dans [Denoyer & Gallinari, 2004]. L'apparition des documents XML, et plus généralement des documents structurés, a modifié la problématique de la RI. Son objectif ne s'arrête plus à la classification des documents pertinents pour une requête d'un utilisateur : un système de recherche d'information doit aussi être capable d'isoler les sous-parties pertinentes d'un document et de les sou-

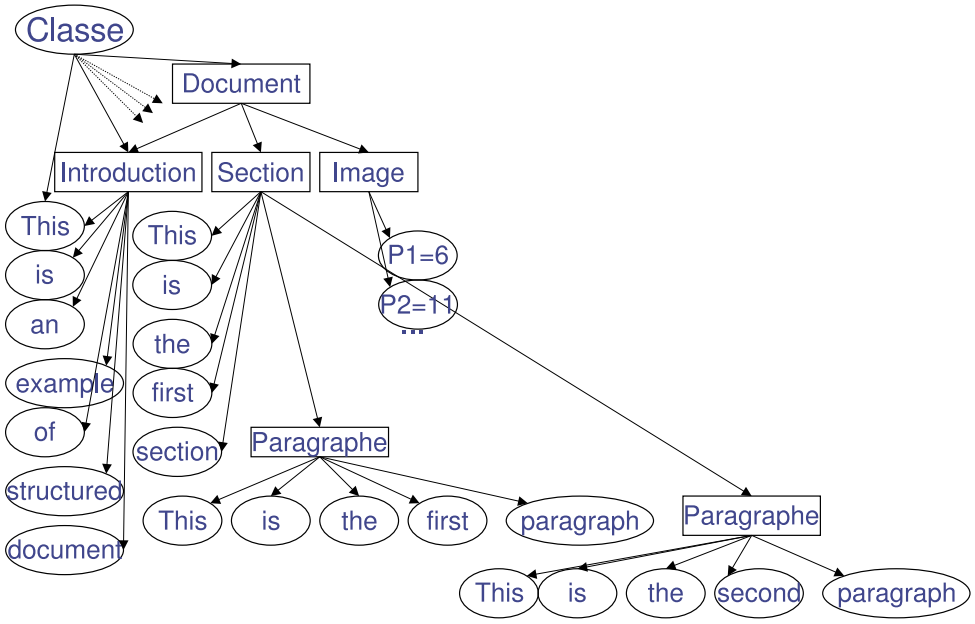

FIG. 4.11 – Exemple d'un réseau bayésien associé à un document structuré

mettre à l'utilisateur. Pour ce faire, le contenu textuel des documents n'est plus le seul élément sur lequel doit reposer la classification, la structure logique du document ainsi que les contenus multimedia doivent être examinés pour apprécier la classe d'une partie ou de la totalité d'un document inconnu. La solution apportée repose sur les RB. Le système construit pour chaque document un réseau propre au document (voir la figure 4.11).

L'auteur distingue artificiellement deux types de nœuds : les nœuds structurels qui marquent les paragraphes, les titres, *etc.* et les nœuds de contenus qui contiennent les informations textuelles, des images, *etc.* Le choix de l'auteur pour la représentation des images est volontairement simple. Une image est un histogramme des couleurs qui la compose. L'image est le seul contenu non-textuel considéré lors des expérimentations. Le système dispose d'une représentation qui intègre la structure logique du document ainsi que ses contenus. Il est alors possible de calculer la classe du document complet en calculant la probabilité de la racine de l'arbre ou seulement la classe d'une partie du document en calculant la probabilité d'un sous-arbre. Trois corpus ont servi pour l'évaluation du système. Le premier corpus est composé d'articles d'informatique aux format XML, le second de pages HTML des sites web de différents départements d'informatique et le dernier de pages HTML variées contenant du texte et des images. Les performances du système ont été comparées avec celles d'un classifieur bayésien naïf, d'un SVM et de deux SVM construits à partir des réseaux bayésiens. Sur les deux premiers corpus où les images sont absentes, les scores des modèles disciminants sont supérieurs aux scores des modèles génératifs, les SVM construits à partir des réseaux bayésiens l'emportant sur le SVM classique. La représentation vectorielle des SVM ne permet pas la représentation des images. Les SVM ne peuvent donc pas classer les documents du dernier corpus. Seuls les résultats des RB avec et sans la représentation sont publiés. La présence des images, même avec une représentation fruste, augmente les performances du système.

Les conclusions de [Denoyer & Gallinari, 2004] mettent en avant l'expressivité du modèle. Toute information dont le contenu à la possibilité d'être traduit par un modèle génératif peut enrichir le modèle et infléchir la décision du système. Les modèles génératifs n'intègrent plus uniquement des informations textuelles mais des informations hétérogènes comme les images, la structure logique d'un document ou encore des données statistiques (voir ci-dessus [Sahami *et al.*, 1998]). Cette richesse dans la description réduit le biais de représentation des objets envisagés par le système lors d'une inférence et améliore ses performances globales.

Ces conclusions sont partagées par [Peshkin & Pfeffer, 2003] dans un travail d'extraction d'information (EI). Le système Bayesian Information Extraction Network (BIEN) extrait automatiquement des informations ciblées telles que la date, le lieu et le (ou les) invités dans un corpus d'annonces de séminaires. Les annonces étant des textes semi-structurés, le système recourt à différentes connaissances linguistiques pour retrouver les informations. Les auteurs énumèrent un ensemble d'informations classiques pour cette tâche (les tokens, les lemmes, les catégories syntaxiques et les entités nommées). Ils tiennent aussi compte des constituants syntaxiques et soumettent l'idée de compléter les EN par des relations d'hyperonymie calculées à partir de WordNet. Ce travail montre comment un modèle plus complexe, le RBD, intègre dans une unique représentation les attributs classiques de l'EI, des indices de surface moins fiables ainsi que l'ordre des composants de l'annonce. Chaque token du texte est représenté par un instant du RBD (voir la figure 4.12).

Le RBD contient les attributs linguistiques observables ainsi que la variable cachée *Cible* dont on recherche l'attribut le plus probable, *Date-debut, Date-fin, Lieu, Invité, Autre* où l'attribut *Autre* décrit les tokens que l'on ne souhaite pas extraire. Une autre variable cachée, *Dernière Cible* conserve l'information du dernier composant rencontré dans la séquence. La variable *Segment du Document* dont les attributs sont *Titre_Document, Corps_Document* situe la position logique du segment dans le document. L'évaluation du RBD a été effectuée sur un corpus de référence pour les annonces de séminaire et les résultats annoncés sont similaires à ceux des meilleurs systèmes de l'état de l'art sauf pour la reconnaissance des attributs *Date-fin* et *Lieu* où le RBD fait mieux que ses concurrents.

4.5.4 Un apprentissage automatique du réseau possible

La dernière étude que nous présentons, [Bouckaert, 2002], est sans doute l'une des plus abouties puisque l'auteur s'assigne une double tâche d'extraction d'information et surtout montre que l'apprentissage automatique du réseau est non seulement réalisable mais aussi bénéfique pour le TAL.

Dans la première tâche d'extraction il lui faut extraire les composants des affiliations d'un corpus d'articles pharmaceutiques et, dans la deuxième tâche, les éléments des références bibliographiques d'un corpus de notices bibliographiques. Les MMC réalisent cette tâche d'extraction comme une tâche de classification et obtiennent généralement de bons résultats. Mais le formalisme est limité car il n'est possible d'observer qu'un seul attribut. L'auteur reproduit l'idée mais avec le modèle plus général des RB. Chaque séquence est découpée en *tokens* d'un ou plusieurs mots. Les tokens sont décrits par des attributs variés. Par exemple, les tokens d'une affiliation sont décrits par les variables :

– *être_institut*, vraie si le token est une des chaînes de l'ensemble {Hospital, University...}, fausse sinon ;

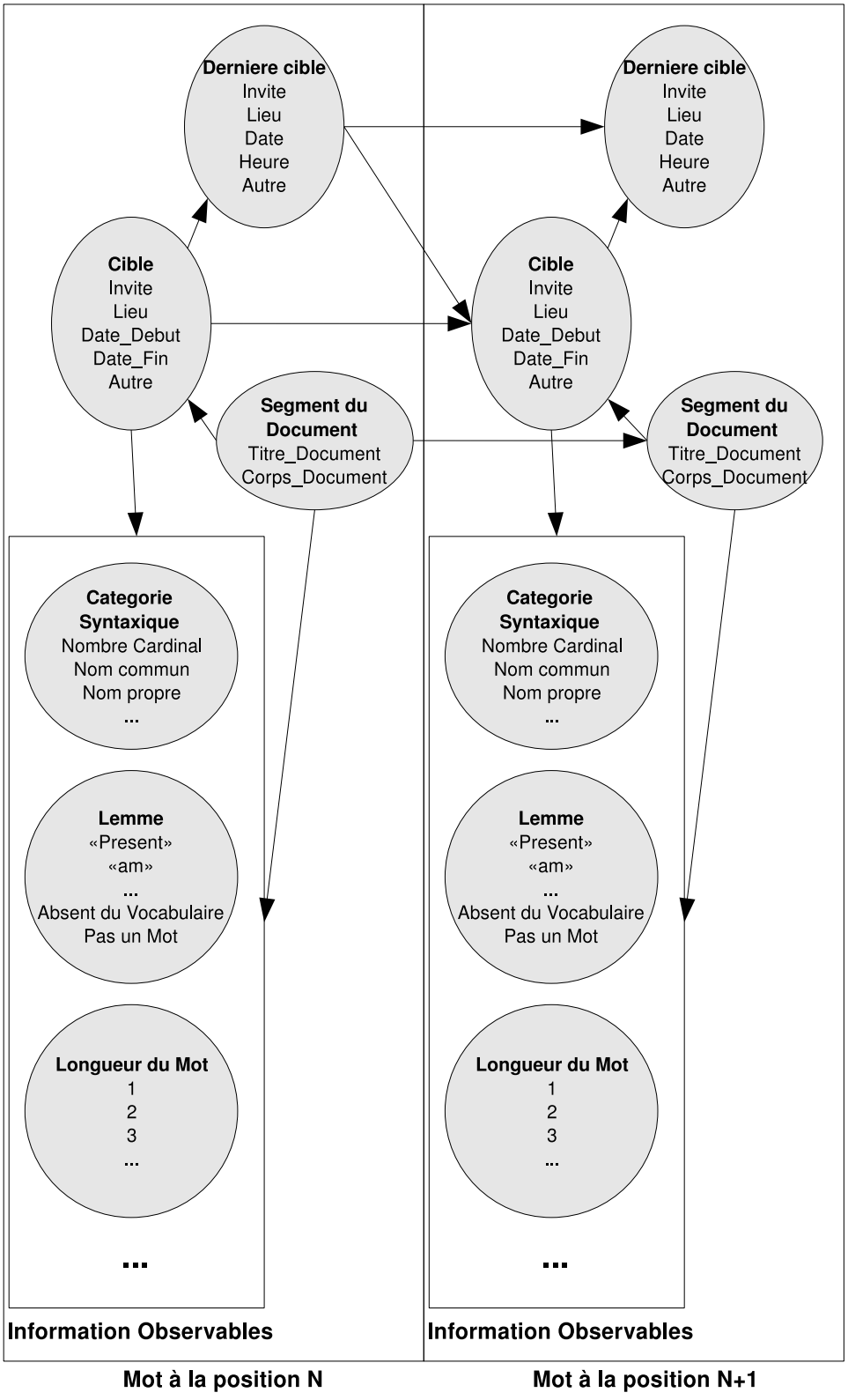

FIG. 4.12 – Un réseau Bayésien Dynamique pour l'extraction automatique d'annonces de séminaires

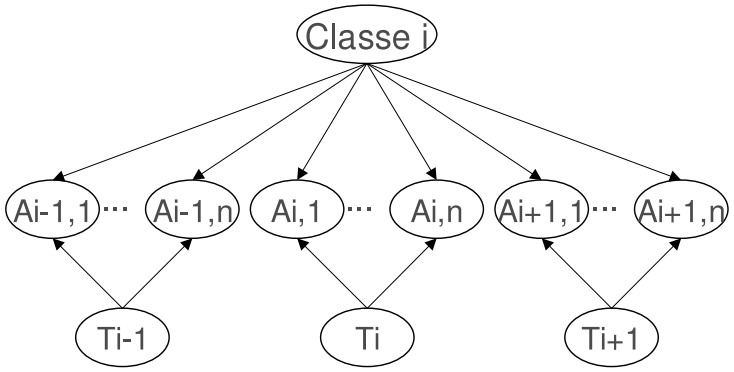

FIG. 4.13 – Un réseau Bayésien pour l'extraction d'information

– *être_code_postal*, vraie si le token est reconnu par une expression régulière ;
– *être_ville*, vraie si le token appartient à un dictionnaire de noms de villes ;
– ...

L'auteur modélise une fenêtre de trois tokens successifs (voir le schéma de l'auteur figure 4.13) et les ordonne grâce aux VA *Ti-1,Ti,Ti+1* dont les valeurs possibles sont *First,Mid,Last*.

L'auteur compare les performances de plusieurs réseaux appris automatiquement. Les paramètres sont appris sur un ensemble de données d'apprentissage complet. Les structures des réseaux sont apprises avec deux algorithmes différents à base de scores, K2 et B, auxquels l'auteur impose différentes structures initiales pour l'apprentissage.

L'évaluation sur la tâche d'extraction d'information des affiliations mesure les performances d'un système à base de règles logiques (87.1% Acc), d'un classifieur bayésien naïf dont les performances se situent autour de 93.5%Acc et de différents RB dont le plus performant obtient un score de 96%. L'ordre des performances est sensiblement identique sur la deuxième tâche avec des performances générales moins bonnes (94.4% pour le meilleur RB). Selon l'auteur, il apparaît que les RB les plus complexes *i.e.* les RB avec la structure la plus profonde, obtiennent les meilleurs résultats bien qu'un plafonnement des résultats se dessine au delà d'une structure admettant plus de 5 générations. Des améliorations sont envisagées avec l'ajout d'attributs spécifiques au domaine et un changement du modèle au profit des RBD pour tenir compte de la classe attribuée au token précédent (une expérience supplémentaire de l'auteur, où la classe des tokens précédant le token à classer est fournie au RB, améliore les résultats de 2%).

4.6 Conclusion

Pour des raisons pratiques, nous avons restreint la présentation de la famille des modèles génératifs aux modèles qui nous intéressent directement dans ce travail. Nous avons défini et détaillé le fonctionnement du modèle d'inférence sur lequel notre système de résolution d'anaphore repose, le modèle des réseaux bayésiens.

Ce modèle est très intuitif puisqu'il représente graphiquement les liens d'influence entretenus entre les différents attributs des objets du monde conçu et permet de quantifier ces influences grâce à un ensemble de probabilités conditionnelles associées au modèle. Il généralise des mo-

dèles d'inférence largement utilisés dans le domaine du TAL, le classifieur bayésien naïf et les modèles de Markov cachés. La structure du réseau souffre cependant de deux contraintes importantes : elle ne doit pas comporter de cycles et une fois fixée elle est définitive. La première contrainte facilite les calculs et ne gêne pas la conception de notre système. La seconde, en revanche, devra être levée dans le chapitre 6 pour modéliser correctement le problème du choix de l'antécédent.

Dans ce travail nous avons traduit le problème de la résolution d'anaphore comme une tâche de classification. Nous aurions pu préférer aux réseaux bayésiens d'autres modèles d'apprentissage qui obtiennent généralement de meilleures performances, les modèles discriminants. La performance est un critère essentiel dans le choix d'un modèle. Mais, pour le domaine du TAL, l'intégration, l'interprétation et l'adaptation des connaissances du modèle nous semblent être des critères plus importants comme le soulignent les travaux de l'état de l'art que nous avons cités.

A l'instar des modèles discriminants, le réseau bayésien tient compte de la fiabilité respective des attributs à sa disposition dans sa décision et offre une étape d'apprentissage pour évaluer l'importance des attributs du corpus traité. Mais ce modèle dispose d'autres avantages par rapport aux modèles discriminants. Nous avons vu dans le chapitre 1 que le biais de représentation est très élevé en TAL en raison d'un choix difficile des attributs discriminants pour une tâche d'annotation particulière. Il est donc nécessaire d'intégrer le plus d'informations possibles au sein du modèle pour réduire ce biais. Le modèle des RB permet d'intégrer facilement des informations de natures variées, alors que cette intégration est plus difficile voire impossible avec les modèles discriminants. Un autre avantage du modèle vient de l'interprétation naturelle des liens d'influence de la structure du réseau, même par des non initiés. Cette propriété conjuguée à la maturité naissante des techniques d'apprentissage automatique de la structure rend possible la découverte de nouveaux liens d'influence pertinents à partir de l'analyse des structures apprises automatiquement. Enfin le choix du modèle des RB n'est pas définitif puisque nous pouvons modéliser un SVM équivalent au RB défini pour une tâche et espérer de meilleurs résultats.

Chapitre 5

Un filtre bayésien pour les pronoms impersonnels

Sommaire

Dans ce chapitre nous présentons un travail préparatoire à la résolution automatique des anaphores. Comme nous l'avons rappelé dans la section 2.1.4, avant toute résolution des anaphores, il convient de distinguer les pronoms impersonnels des pronoms anaphoriques. Nous proposons ici un filtre pour les pronoms impersonnels reposant sur un réseau bayésien. Notre intention est de valider nos choix et d'évaluer les qualités de notre modèle sur cette tâche de classification qui est plus simple que celle du choix de l'antécédent.

5.1 Comment filtrer les pronoms impersonnels ?

Dans cette section nous cherchons uniquement à différencier les pronoms impersonnels des pronoms anaphoriques.

Certains auteurs [Boyd *et al.*, 2005], [Evans, 2001], suggèrent un classement plus riche pour le pronom *it*. Après une étude en corpus [Evans, 2001] distingue sept types de pronoms différents que nous reportons brièvement avec les exemples de l'auteur :

- les pronoms anaphoriques nominaux : le pronom renvoie à un syntagme nominal précédemment mentionné dans le discours,
  Do not sweep [the dust]$_i$ when dry, you will only recirculate [it]$_i$.
- les pronoms anaphoriques clausaux : le pronom se rapporte à une proposition précédente,
  [One day in 1970, fifty thousand women marched down fifth Avenue in New York]$_i$. [It]$_i$ is said to have been the biggest women's gathering since suffrage days.
- les pronoms anaphoriques d'action : le pronom réfère à un syntagme verbal d'une proposition précédente,
  Mays [walloped four home runs in a span of nine innings]$_i$. Incidentally, only two did [it]$_i$ before a home audience.
- les pronoms cataphoriques : le pronom réfère toujours à un antécédent qui est un syntagme nominal, mais il le précède,
  when [it]$_i$ fell, [the glass]$_i$ broke.
- les pronoms discursifs[27] : le pronom est anaphorique mais le lecteur doit recourir à des inférences non triviales pour déterminer le référent du pronom qui peut être un thème abstrait du discours,
  Always use a tool for the job it was designed to do. Always use tools correctly. If [it] feels very awkward, stop
- les pronoms impersonnels : le pronom *it* ne renvoie à aucun objet du discours, il est sémantiquement vide ou encore impersonnel,
  It is worth having more than one size or a good-quality set with interchangeable bits.
- le pronom idiomatique ou stéréotypique : le pronom est impersonnel et il apparaît dans des expressions figées de la langue,
  I take it you're going now

L'auteur analyse finement ce phénomène car il entend spécialiser l'apprentissage des règles de reconnaissance pour chaque type de pronom. Cette granularité est trop précise pour la tâche que nous nous assignons dans ce chapitre. Nous cherchons uniquement à distinguer les emplois impersonnels des emplois anaphoriques pour lesquels il convient de rechercher l'antécédent (voir la section 5.1.3 pour une définition formelle de cette tâche). Nous regroupons dans la catégorie impersonnelle les pronoms idiomatiques et les pronoms impersonnels, les autres catégories sont considérées comme anaphoriques.

5.1.1 Les filtres reposant sur des informations syntaxiques

L'un des premiers systèmes distinguant les pronoms *it* impersonnels et anaphoriques a été conçu par [Paice & Husk, 1987]. Les auteurs s'appuient sur un ensemble de règles de logique du premier ordre pour distinguer les occurrences impersonnelles du pronom. Si l'une de ces règles reconnaît une séquence qui contient une occurrence du pronom *it*, le système conclut que le pronom est impersonnel. Si aucune règle ne reconnaît la séquence, le système conclut que le pronom est anaphorique. Notons que cet ensemble de règles peut facilement être réexprimé par un vecteur d'attributs numériques. [Paice & Husk, 1987] remarquent que les séquences qui introduisent les *it* impersonnels partagent une forme remarquable (elles commencent par un *it* et se terminent par un délimiteur comme *to, that, whether...*) même si le détail des règles va-

[27] Textuellement, *Discourse topic*.

rie d'un délimiteur à l'autre. Les contraintes portent sur le contexte gauche (le pronom ne doit pas être précédé immédiatement par certains mots comme *before, from, to...*), sur le nombre de mots séparant le pronom et le délimiteur (ce nombre doit être inférieur à 25 mots), et enfin sur les éléments lexicaux présents entre le pronom et le délimiteur (la séquence doit contenir ou ne pas contenir certains mots appartenant à des classes connues de mots, par exemple les mots exprimant une modalité propositionnelle, (*e.g. certain, known, unclear...*). Les tests réalisés par Paice montrent que ces règles réalisent un bon score avec 91.4%*Acc* sur un corpus technique. Cependant leurs performances se dégradent si on les applique à des corpus de natures différentes. On observe que le nombre de Faux Positifs augmente. Cette augmentation est dûe à une erreur d'estimation, les attributs qui sont discriminants sur les corpus techniques ne l'étant plus dès lors que la nature du corpus change. Nous commentons en détail ce phénomène lors de la dernière expérience de cette section 5.3.3.

Afin d'éviter cet écueil, [Lappin & Leass, 1994] propose des règles plus contraintes sous la forme d'automates à états finis. Les automates reconnaissent des séquences spécifiques comme *It is not/may be<Modaladj>* ; *It is <Cogv-ed> that <Subject>* où <Subject> est un syntagme nominal qui doit être le sujet d'un verbe de la phrase, *<Modaladj>* et *<Cogv>* sont des classes d'adjectifs modaux et de verbes cognitifs connus pour introduire des *it* impersonnels (par exemple *necessary, possible* et *recommend, think*). Ce système a une bonne précision (il produit peu de Faux Positifs), mais il a un mauvais rappel (il produit beaucoup de Faux Négatifs). Avant d'expliquer cette chute du rappel remarquons que, comme les automates décrivent entièrement un ensemble de séquences finies, elles-mêmes de longueurs finies et connues, il est possible de retraduire ces automates par des vecteurs d'attributs numériques mais ceux-ci sont plus complexes que dans les vecteurs d'attributs codant les règles logiques de [Paice & Husk, 1987]. Il faut noter également que les propriétés '*suivre_immédiatement*' et '*appartenir_à_ la_ classe_sémantique*' introduites par [Lappin & Leass, 1994] et non présentes dans [Paice & Husk, 1987] ne sont pas des attributs fiables. Seules les séquences exactes sont reconnues, ce qui pose le problème des petites variations qu'on observe en corpus. Il est par ailleurs difficile d'obtenir des classes d'adjectifs et de verbes exhaustives. Voici deux phrases tirées de notre corpus, qui ne sont pas reconnues par les automates de [Lappin & Leass, 1994] : *It is well documented that treatment of serum-grown...* et *It is generally accepted that Bcl-2 exerts...* Les occurrences du pronom ne sont pas considérées comme impersonnelles car le verbe *documented* ne figure pas dans la classe *<Cogv>* et parce que l'adverbe *generally* n'apparaît pas dans l'automate précédent.

5.1.2 Les filtres s'appuyant sur des informations distributionnelles

Devant le bruit produit par les attributs de [Lappin & Leass, 1994], [Evans, 2001] renonce à exploiter des propriétés aussi complexes et se concentre sur des attributs fiables. Il exploite uniquement des indices de surface. Une étape d'apprentissage réduit l'erreur d'estimation en déterminant le poids des attributs discriminants pour le domaine du corpus. Evans considère 35 attributs, des indices de surface syntaxiques et contextuels (ex. la position du pronom dans la phrase, le lemme du verbe suivant...) qui sont extraits d'un ensemble d'exemples annotés manuellement. Le système classe les nouvelles instances au moyen d'une métrique des *k* plus proches voisins. Les premiers essais réalisent un score de 71.31%*Acc* satisfaisant sur un corpus de langue générale. [Litran *et al.*, 2004] reproduit un essai identique sur un corpus de Géno-

mique. Parmi les attributs de [Evans, 2001], il se limite aux 21 plus pertinents et classe les nouvelles instances avec une Machine à Support de Vecteur (SVM). La SVM obtient un score de 92.71%*Acc*.

Ces deux derniers systèmes d'apprentissage reposent uniquement sur des indices de surface. Pour se soustraire au problème du bruit des attributs du système de [Lappin & Leass, 1994] et de l'erreur d'estimation de celui de [Paice & Husk, 1987], ils délaissent les attributs utilisés par ces derniers mais ce choix est critiquable puisqu'il augmente le biais de représentation déjà élevé pour le TAL. En dépit de leur manque de fiabilité, les attributs qui traduisent des connaissances linguistiques sont pertinents pour notre tâche. Nous soutenons que ces attributs sont importants et doivent intégrés au système en dépit de leur manque de fiabilité.

5.1.3 La distinction des pronoms, un problème de classification

La distinction des pronoms impersonnels des pronoms anaphoriques se ramène, comme un grand nombre de tâche du TAL, à un problème d'étiquetage d'une séquence du texte, ici le pronom. Attribuer une étiquette revient à associer à la séquence la classe la plus probable. Définissons formellement cette tâche de classification des pronoms impersonnels :

Soit,

- *Corpus* un ensemble de textes d'un même domaine,
- *Corpus_entrainement* et *Corpus_test* deux sous-ensembles stricts disjoints de *Corpus*,
- C_1 et C_2 les classes des occurrences des pronoms impersonnels et anaphoriques présents dans *Corpus*.

On définit e une occurrence d'un pronom présent dans *Corpus_entraînement*. Le pronom est décrit par un vecteur $v = a_1, ...a_v$ d'attributs à valeurs dans un ensemble fini quelconque où chaque attribut a_i a été sélectionné sur la base d'une analyse humaine de *Corpus_entrainement*. Pour tout e de *Corpus_test*, notre classifieur attribue

la classe C_1 à l'exemple e si P(Pronom=Impersonnel|e)$\geq$P(Pronom=Anaphorique|e) et

la classe C_2 sinon.

Ni les règles ni les indices de surface ne sont des indicateurs fiables du statut du pronom mais ils sont complémentaires : comme ils produisent des faux positifs et des faux négatifs différents, il est intéressant de les combiner. Les règles de [Lappin & Leass, 1994] étiquètent avec précision mais s'appliquent à très peu d'occurrences de pronoms. Une solution consiste à compléter leur faible rappel par celui, très large, des règles de [Evans, 2001]. Seulement ces règles sont imprécises et doivent être corroborées par la présence des indices de surface de [Litran *et al.*, 2004] et de [Evans, 2001]. Le classifieur basé sur un réseau bayésien que nous présentons permet, en exploitant des liens de dépendance, de combiner ces indicateurs entre eux et de tenir compte de leurs fiabilités respectives. Il fournit un résultat global supérieur à celui obtenu à partir de chaque indicateur exploité isolément.

5.2 Un classifieur bayésien pour la reconnaissance des pronoms impersonnels

5.2.1 Modélisation et emploi du classifieur bayésien

Un choix de représentation aurait pu être de réexprimer les attributs encodés dans les règles des systèmes de [Paice & Husk, 1987] et de [Lappin & Leass, 1994] et de les ajouter à ceux des méthodes par apprentissage en un unique vecteur d'attributs cohérent. Mais cette approche est difficile à mettre en œuvre : il serait nécessaire pour chaque nouvelle méthode de réexprimer les attributs qu'elle exploite et de repenser la cohérence du nouveau vecteur d'attributs. Notre approche est plus simple : nous considérons les règles de [Paice & Husk, 1987] et de [Lappin & Leass, 1994] comme des attributs à part entière que nous complétons par les attributs de [Litran *et al.*, 2004] et de [Evans, 2001] qui ne sont pas déjà inclus dans les règles précédentes.

Le classifieur bayésien (CB) que nous avons utilisé pour classer les occurrences impersonnelles est décrit par la figure 5.1. L'attribut représentant le fait qu'une règle de [Lappin & Leass, 1994] ait reconnu une séquence est coloré en gris, en blanc celui qui correspond aux règles de [Paice & Husk, 1987], enfin en noir les attributs de [Litran *et al.*, 2004] et [Evans, 2001]. Le nœud de prédiction est le nœud *Pronom*, au centre. Il estime la probabilité pour une occurrence donnée du pronom d'être impersonnelle ou anaphorique.

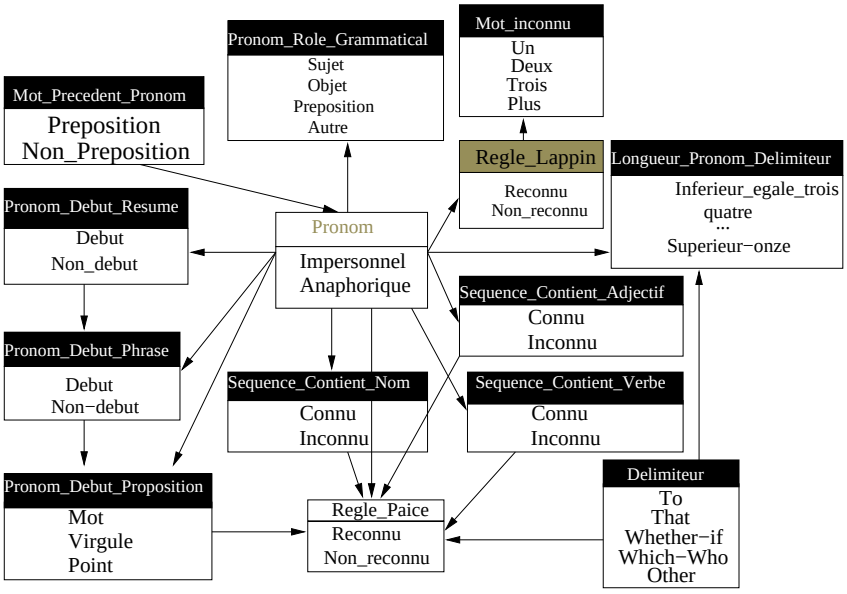

FIG. 5.1 – Un réseau bayésien pour la classification des pronoms *it* impersonnels

A l'heure actuelle, seule une expertise linguistique des attributs décrits dans l'état de l'art 5.1 justifie les attributs retenus et la structure de notre CB. Nous envisageons des expériences supplémentaires à celles que nous présentons dans cette thèse. La comparaison de notre structure avec des structures apprises automatiquement et exhibant des dépendances significatives ignorées lors de notre expertise permettrait de vérifier et de modifier la structure du CB actuelle.

Nous donnons la signification qualitative des variables aléatoires (VA) et des différents liens d'influence qui composent notre RB.

[`Mot_Precedent_Pronom`] Si le mot qui précède immédiatement le pronom est une préposition, le pronom est sans aucun doute anaphorique. C'est le seul attribut qui permet de discriminer sans erreur les occurrences anaphoriques de pronom, d'où l'arc de *(Mot_Precedent_Pronom)* pointant sur *Pronom*.

[`Pronom_Debut_Proposition`] La position du pronom dans la proposition influence la probabilité que le pronom soit impersonnel. S'il est précédé d'un point ou d'une virgule cela renforce la probabilité que le pronom soit impersonnel. Il a plus de chance d'être anaphorique s'il est précédé d'un mot.

[`Pronom_Debut_Phrase` et `Pronom_Debut_Resume`] Si le pronom est l'un des 3 premiers mots de la phrase ou du résumé, nous considérons que le pronom commence une phrase (resp. un résumé). Cela renforce la probabilité qu'il soit impersonnel. Si le pronom est en début de résumé alors il est aussi en début de phrase. Et le fait d'être en début de phrase implique que le pronom est précédé d'un point.

[`Pronom_Role_Grammatical`] La probabilité que le pronom soit impersonnel augmente s'il est sujet et baisse dans les autres cas.

[`Regle_Lappin`] Si la séquence qui contient le pronom est reconnue par un automate de [Lappin & Leass, 1994], la probabilité qu'il soit impersonnel augmente fortement.

[`Mot_Inconnu`] Nous avons modifié les automates de [Lappin & Leass, 1994] pour qu'ils puissent admettre un maximum de 3 mots inconnus dans les séquences qu'ils reconnaissent. Plus il y a de mots inconnus dans la séquence, moins l'automate est fiable.

[`Delimiteur`] Cette VA fixe la nature du premier délimiteur rencontré dans la séquence.

[`Longueur_Pronom_Delimiteur`] Cette VA détermine le nombre de mots qui séparent le pronom du délimiteur. Nous avons restreint la longueur maximale à 10 mots[28]. Plus la distance entre le pronom et le délimiteur est importante moins le délimiteur est fiable. Cette dépendance est exprimée par l'arc allant de `Delimiteur` à `Longueur_Pronom_Delimiteur`.

[`Sequence_Contient_{nom,verbe,adjectif}`] Ces VA déterminent si, parmi les mots présents entre le pronom et le délimiteur, figure un nom (un verbe, un adjectif ou un adverbe) qui a déjà été rencontré entre une occurrence de pronom impersonnel et un délimiteur dans une ou plusieurs séquences du corpus d'entraînement. La présence d'un tel mot augmente la probabilité pour le pronom d'être impersonnel. Nous avons acquis automatiquement ces classes de mots (la procédure est commentée dans la section 5.2.2)

[`Regle_Paice`] Si la séquence qui contient le pronom est reconnue par une règle de [Paice & Husk, 1987], la probabilité qu'il soit impersonnel augmente. Les arcs indiquent que l'importance de cette influence dépend de la nature du délimiteur, de la position du pronom dans la proposition et de la présence d'un mot connu entre le délimiteur et le pronom.

Les valeurs des probabilités conditionnelles sont apprises automatiquement avec l'approche du maximum de vraisemblance à partir des données d'un corpus d'entraînement. Aucune valeur

[28] Une limite établie à partir des observations faites sur nos corpus.

n'est manquante à l'exception de la variable *Pronom_Role_Grammatical*. L'analyse syntaxique ayant échoué sur certaines phrases[29], le rôle grammatical du pronom est indisponible. Pour ces expériences préliminaires, nous avons choisi une solution facile à mettre en œuvre. Nous avons retranché 20% des pronoms du corpus (soit environ 700 pronoms) que nous avons analysé automatiquement, corrigé et complété [Alphonse *et al.*, 2004]. La valeur des probabilités conditionnelles calculées sur cette partie de corpus nous servant de valeur *a priori* pour la VA *Pronom_Role_Grammatical*.

Illustrons avec un exemple concret. Parmi 2 000 occurrences du pronom *it* d'un corpus d'entraînement, les règles de [Lappin & Leass, 1994] reconnaissent 649 des 727 pronoms impersonnels de ce corpus d'entraînement et elles reconnaissent à tort 17 pronoms anaphoriques. Nous calculons les probabilités du nœud *Regle_Lappin* :

P(Regle_Lappin=Reconnu|Pronom=Impersonnel)=0.892 et

P(Regle_Lappin=Reconnu|Pronom=Anaphorique)=0.013[30],

qui expriment les valeurs attendues pour les FN et les FP produits par les règles de [Lappin & Leass, 1994].

Durant l'étape d'inférence, nous appliquons toutes les règles précédentes et nous déterminons les attributs restants sur les séquences contenant un *it* puis nous assignons les valeurs des observations aux VA correspondantes. Une nouvelle probabilité est calculée pour la variable du nœud `Pronom` : si elle est supérieure ou égale à 50 % l'occurrence du pronom est classée impersonnelle ; elle est anaphorique dans le cas contraire. Considérons la phrase extraite d'un corpus de Génomique :

It had previously been thought that ZEBRA's capacity to disrupt EBV latency...

et appliquons les règles de [Lappin & Leass, 1994] et les règles de [Paice & Husk, 1987] sur cette séquence :

- comme aucune règle de [Lappin & Leass, 1994] ne reconnaît la séquence, même en tolérant 3 mots inconnus, nous posons les observations *P(Regle_Lappin=Non-reconnu)=1* et *P(Mots_inconnus=Plus)=1*,
- comme une règle de [Paice & Husk, 1987] reconnaît la séquence avec 4 mots entre le pronom et le délimiteur *that*, nous posons *P(Regle_Paice=Reconnu)=1*, *P(Longueur_Pronom_Delimiteur=4)=1* et *P(Delimiteur=That)=1*,
- parmi les attributs restants, nous vérifions différents facteurs :
  - la séquence commence une phrase,
  - cette phrase n'est pas la première du résumé,
  - la séquence contient l'adverbe *previously* et le verbe *think*, des mots qui figurent par ailleurs entre des pronoms et les délimiteurs dans le corpus d'entraînement,
  - le rôle grammatical du pronom étant correctement calculé, nous posons *P(Pronom_Role_Grammatical=Sujet)=1* (lorsque l'analyse échoue, nous laissons les valeurs *a priori*)[31].

La probabilité *a priori* pour le pronom d'être impersonnel est de 35.5%[32] mais les nœuds du classifieur sont modifiés pour tenir compte des observations précédentes. La probabilité *a*

[29] Voir la section 5.2.2 pour des précisions supplémentaires sur l'analyseur syntaxique utilisé.

[30] Soit *P(Regle_Lappin=Reconnu|Pronom=Impersonnel)=649/727=0.8927* et *P(Regle_Lappin=Reconnu|Pronom=Anaphorique)=17/(2000-727)*

[31] Les autres valeurs des nœuds sont assignées en conséquence.

[32] Nous avons déterminé la probabilité P(Pronom=Impersonnel) avec le calcul exacte décrit dans 4.1.2.

posteriori calculée pour cette occurrence est égale à 99.7%[33] et le classifieur la considère comme impersonnelle.

5.2.2 Présentation des corpus, des systèmes et du protocole expérimental

Les corpus d'expérience

Corpus Génétique : Le premier corpus avec lequel nous avons travaillé est un corpus de génomique [Alphonse *et al.*, 2004]. Il a été construit à partir de la base *Medline*[34], une base documentaire spécialisée dans les articles de recherche de génomique, qui a été interrogée avec les mots clés *bacillus subtilis, transcription factors, Human, blood cells, gene and fusion*. Nous en avons extrait 11 966 résumés (environ 5 millions de mots) d'où nous avons isolé 3 206 phrases contenant 3 347 occurrences du pronom *it*.

Corpus Management : Le second corpus, le corpus Management, est un corpus que nous avons construit spécialement pour nos expériences. Il est composé à partir d'articles de recherche en sciences économiques et sociales de la base de connaissances *Econlit*[35]. Nous avons sélectionné les 1 000 premiers résumés (environ 225 000 mots) retournés en réponse à la requête *benefits and market*. Ces résumés contiennent 585 occurrences du pronom *it*.

Annotation des corpus : Deux annotateurs humains ont analysé l'ensemble des occurrences du pronom *it*. Ils ont classé chaque occurrence soit comme anaphorique soit comme impersonnelle. Les annotateurs furent, après discussion, d'accord sur toutes les étiquettes attribuées aux occurrences.

Architecture générale du système

Nous avons constitué automatiquement les classes des noms, verbes et adjectifs nécessaires aux variables `Sequence_Contient_{nom,verbe,adjectif}` et nous avons enrichi les automates de [Paice & Husk, 1987]. Nous nous sommes servi d'une constatation rudimentaire : nous avons considéré que les occurrences de pronoms qui se trouvent dans les trois premiers mots de la première phrase (autre que le titre) d'un document sont impersonnelles. Nous avons interrogé le moteur de recherche Medline avec le terme "blood" et travaillé sur les 3 premiers gigaoctet de résumés retournés. Parmi les 13 517 premières phrases contenant le pronom en première position, nous avons écarté 273 phrases qui ne présentaient pas de délimiteur dans les 13 premiers mots, *e.g.*

it was investigated the influence of a diet with zinc supplementation on dynamic of glycaemia, lipid profile, blood pressure and weight in type 2[...]

[33] Nous avons là encore déterminé la probabilité *P(Pronom=Impersonnel|Regle_Lappin=Non-reconnu, Mots_inconnus=Plus, Regle_Paice=Reconnu, Longueur_Pronom_Delimiteur=4, Delimiteur=That, Mot_Precedent_Pronom=Non_Preposition, Pronom_Debut_Proposition=Point, Pronom_Debut_Phrase=Non_debut, Pronom_Debut_Resume=Non_debut, Pronom_Role_Grammatical=Sujet, Sequence_Contient_nom=Inconnu, Sequence_Contient_verbe=Connu, Sequence_Contient_adjectif=Connu)* avec le calcul exacte décrit dans 4.1.2.

[34] Cette base de donnée est accessible sur : `http://www.ncbi.nlm.nih.gov/entrez/`

[35] Cette base de donnée est accessible sur : `http://www.econlit.org/`

Nous avons obtenu 2177 séquences *it...délimiteur* différentes qui complètent les automates proposés par [Paice & Husk, 1987]. Nous avons enfin regroupé les noms, verbes, adjectifs et adverbes étiquetés par le TreeTagger présents dans les séquences.

Pour déterminer les valeurs des autres attributs nous avons employé une série de transducteurs à états finis grâce au logiciel *Unitex*[36]. Nous avons extrait le rôle syntaxique du pronom au moyen du *Link Parser* pour le corpus Management et d'une version du *Link Parser*, nommée *BioLG*, adaptée au domaine de la génomique [Aubin *et al.*, 2005] pour le corpus Génétique. Nous avons écrit et interrogé notre réseau bayésien avec l'API de *Netica* (en langage C). Les caractéristiques techniques de l'outil sont commentées en annexe de [Naïm *et al.*, 2004][37]. Un ensemble de scripts perl permet d'interfacer les entrées et les sorties des différents composants ainsi que la gestion, via des requètes SQL, d'une base de données permettant l'apprentissage des fréquences et l'enregistrement des résultats. La figure 5.2 schématise le système complet.

Les systèmes en compétition

Afin d'évaluer notre approche, nous avons mis en concurrence 4 versions différentes de notre système ainsi que le SVM de [Litran *et al.*, 2004]. Les deux premiers systèmes reposent uniquement sur les règles de [Lappin & Leass, 1994] et [Paice & Husk, 1987] que nous avons réimplémentées. Les performances de ces systèmes sont celles de nos transducteurs appliqués sur notre corpus. Notre classifieur bayésien et le classifieur bayésien naïf qui lui est associé (voir la section 4.3.1 et la figure5.3) sont les troisièmes et quatrièmes systèmes. Ils sont tous deux paramétrés à partir des fréquences d'un corpus d'entraînement.

Le dernier système est le SVM utilisé par [Litran *et al.*, 2004]. Cependant, nous n'avons pas appliqué ce système directement sur nos données. Nous avons ajouté les scores, qui ont été obtenus par ce système sur un corpus de génomique similaire, pour les comparer aux nôtres[38]. Les attributs utilisés par le SVM de [Litran *et al.*, 2004] ne sont que des indices de surface, nous prévoyons par conséquent de compléter ces résultats par ceux d'un SVM réalisant sa classification à partir des mêmes données que celles que nous avons soumises au CB. Comme [Denoyer, 2004] (voir la section 4.4), nous espérons de meilleures performances avec cette méthode de classification.

Le protocole d'expérimentation

Nos corpus étant de taille moyenne, nous avons procédé à une validation croisée pour valider les résultats de nos expériences. Nous sélectionnons aléatoirement deux tiers d'un de nos corpus pour calculer les probabilités conditionnelles *a priori*. Nous appliquons ensuite notre CB, ainsi que le classifieur bayésien naïf (CBN) qui lui est associé, paramétrés grâce à ces probabilités sur le tiers restant ainsi que les systèmes à base de règles. Nous réitérons 20 fois ces opérations pour obtenir la moyenne des performances de chaque système sur le corpus choisi.

[36] Ce logiciel est téléchargeable à l'adresse : `http://www-igm.univ-mlv.fr/~unitex/`.
[37] Une version d'évaluation complète est disponible à l'adresse `http://www.norsys.com/`.
[38] Les valeurs FP et les FN n'ont pas été publiées.

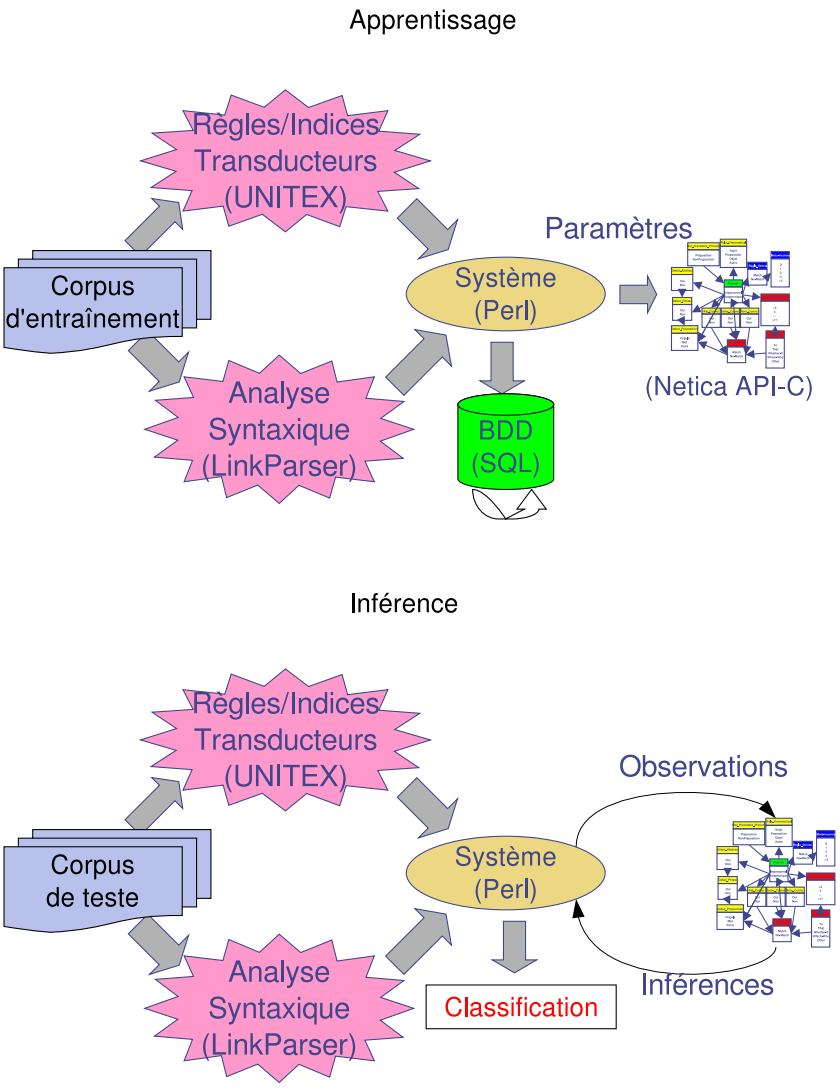

FIG. 5.2 – Architecture du système pour la classification des pronoms *it* impersonnels

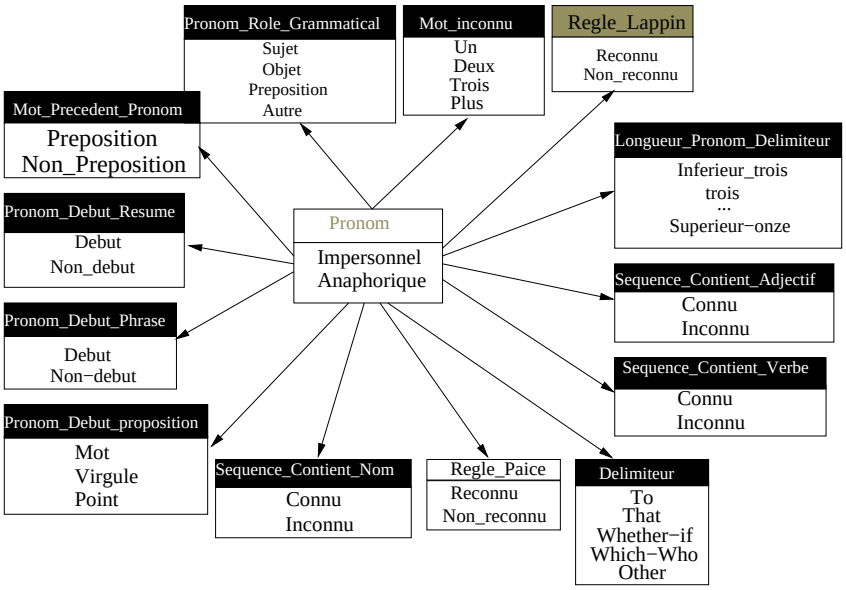

FIG. 5.3 – Un classifieur bayésien naïf pour la classification des pronoms *it* impersonnels

Méthode	Résultats		
Règles de [Lappin & Leass, 1994][39]	88.11 %	12.8	169.1
Règles de [Paice & Husk, 1987][40]	88.88 %	123.6	24.2
SVM de [Litran *et al.*, 2004]	92.71 %	-	-
Classifieur Bayésien naïf	92.58 %	74.1	19.5
Classifieur Bayésien	**95.91 %**	**21.0**	**38.2**

TAB. 5.1 – Résultats des prédictions pour la classification des pronoms impersonnels (Exactitude/Faux Positifs/Faux Négatifs)

5.3 Analyse des résultats

5.3.1 La classification bayésienne, une stratégie performante pour le TAL

L'objectif de cette première expérience est de mesurer les performances de notre CB et de les comparer avec celles des systèmes de l'état de l'art. Nous avons effectué ces mesures sur le corpus de Génétique. Le tableau 5.1 résume les moyennes des résultats en exactitude (voir 2.1.5) obtenus par les méthodes de l'état de l'art décrites précédemment et celles de nos classifieurs. Ces résultats montrent que le CB produit une meilleure classification que les autres systèmes, notamment les systèmes à base de règles.

Ces résultats valident notre approche : la reformulation du problème de la distinction des pronoms impersonnels et anaphoriques en un problème de classification bayésienne permet d'exploiter tous les attributs pertinents et de corriger le bruit d'un attribut par la fiabilité des autres attributs. Les systèmes à base de règles sont entièrement assujettis à la fiabilité des attributs et leurs résultats confirment les craintes soulevées dans la section 5.1 : un faible rappel

Délimiteur	Répartition des FP selon leurs délimiteurs sur l'ensemble des itérations	Nombre de FP pour la $1^{ère}$ itération	Causes
to	47%	10	2 Erreurs d'annotation 5 Transducteurs imprécis 7 Descriptions non discriminantes 5 Configurations spécifiques
that	17%	1	
whether_if	17%	3	
who_which	19%	5	

TAB. 5.2 – Analyse des Faux Positifs produits par le CB

pour les règles de [Lappin & Leass, 1994] et une mauvaise précision pour celles de [Paice & Husk, 1987].

Nous avons étudié les erreurs de notre système en nous focalisant sur la première itération de la validation croisée. Rappelons brièvement les définitions des faux positifs (FP) et des faux négatifs (FN) : un FP correspond à une occurrence d'un pronom anaphorique étiquetée comme impersonnelle par le système et un FN correpsond à une occurrence d'un pronom impersonnel étiquetée comme anaphorique. Nous avons écarté les FP et les FN qui dépendent du paramétrage grâce à une heuristique simple. Selon le tirage d'une itération, une phrase donnée peut appartenir soit au corpus d'entraînement soit au corpus de test. Si pour toutes les itérations où cette phrase apparaît dans le corpus de test le système se trompe sur la classe d'un pronom de cette phrase, nous pouvons supposer que cette conclusion du système est indépendante de la variation des paramètres. Cette distinction nous permet de déterminer les erreurs invariables du système.

Analyse détaillée des FP

Notre CB, qui a une bonne précision, étiquette néanmoins des pronoms impersonnels alors qu'ils ne le sont pas. Les 19 FP de la première itération ont été reconnus par une règle de [Paice & Husk, 1987]. Le tableau 5.2 les classe selon leur délimiteur et résume les causes des erreurs du système (nous donnons les phrases associées en annexe 1 dans le tableau 1).

2 erreurs du système sont dues à une erreur d'annotation dans le corpus annoté. Pour ces deux pronoms les valeurs réelles sont impersonnelles, les valeurs calculées par le système sont impersonnelles mais les valeurs dans le corpus annoté, qui sont fausses, sont anaphoriques. Lors du calcul de ses performances le système compare les valeurs qu'il a calculées, impersonnelles, avec les valeurs fausses du corpus annoté, anaphoriques. Il retourne donc deux FP qui ne sont pas des faux positifs. Ces deux pronoms ont fait l'objet d'un désaccord entre les annotateurs et leurs étiquettes n'ont pas été corrigées dans le corpus annoté.

5 erreurs sont dues à l'imprécision des transducteurs que nous avons implémentés. Les valeurs réelles de ces 5 pronoms sont anaphoriques mais les valeurs calculées par le système sont impersonnelles. Le premier pronom est accepté par l'automate *it [is] likely that*, il est donc classé impersonnel. L'automate permet l'absence de la copule *is* et doit être corrigé. Lorsque la copule est présente la séquence introduit un pronom impersonnel comme dans l'exemple :

> *It is likely that a putative promoter sequence may be present within this 72 nucleotides region for the expression of this cry gene in E. coli.*

Lorsqu'elle est absente, la séquence introduit un pronom anaphorique comme dans l'exemple[41] :

> *The results of electrophoretic mobility shift assays using DNA with [5'-(CGG)(n)-3']$_1$ repeats of different lengths render [it]$_1$ likely that oligomers of the p20 protein bind to the repeat.*

Les quatre autres erreurs sont dues à la reconnaissance d'une séquence plus longue que celle qui était attendue. Par exemple un transducteur reconnaît comme délimiteur de la séquence suivante la deuxième occurence de la préposition *to* et ignore la première :

> *It has been shown [to]$_1$ possess anti-inflammatory activity in a variety of animal models and more recently [to]$_2$ inhibit...*

La séquence reconnue, qui est plus longue que celle attendue, contient un adjectif/adverbe, un nom et un verbe connu ce qui augmente la probabilité à 99.9% et le système classe le pronom impersonnel alors qu'il est anaphorique. Si le transducteur avait reconnu la séquence attendue (*i.e It has been shown to*) la probabilité pour le pronom d'être impersonnel aurait été égale à 0.058%.

7 erreurs sont dues à une description non discriminante dans notre représentation de certaines séquences. Considérons par exemple les séquences :

1. *However, it was very homologous to the promoter...*
2. *It is clinically useful to monitor the cytogenetic...*

Le pronom de la première séquence est anaphorique et le pronom de la seconde séquence est impersonnel. Le système classe les deux pronoms impersonnels. Pour comprendre pourquoi le système se trompe sur le pronom de la première séquence, il faut regarder les vecteurs décrivant ces deux séquences. Soient v_1 le vecteur décrivant la première séquence et v_2 le vecteur décrivant la seconde séquence, on constate que :

v_1=v_2={`Pronom_Regle_Grammatical=Sujet, Mot_Inconnu=Plus, Regle_Lappin=Non_reconnu, Longueur_Pronom_Delimiteur=3, Sequence_Contient_Adjectif=Connu, Sequence_Contient_Verbe= Inconnu, Sequence_Contient_Nom=Inconnu, Delimiteur=To, Regle_Paice=Reconnu, Pronom_Debut_Proposition=Point, Pronom_Debut_Phrase=Debut, Pronom_Debut_Resume=Non-debut, Mot_Precedent_Pronom=Non-Preposition`}

Les deux vecteurs sont égaux et la probabilité associée à ces vecteurs est égale à 56.3%. Sans ajouter de nouveaux attributs, il est impossible pour notre système de distinguer ces deux séquences et de classer correctement la première.

Les 5 dernières erreurs du système sont dues à une configuration spécifique : la conjonction d'une longueur importante de la séquence et de la présence de plusieurs mots connus. Considérons en exemple la séquence :

> *...cell transplantation (PBSCT), whereas it could not be detected in four patients who had been...*

[41] Nous avons demandé la contre-expertise d'un biologiste pour retrouver l'antécédent de l'anaphore dans cette phrase qui est difficile a comprendre pour un non-spécialiste du domaine.

Délimiteur	Répartition des FN selon leurs délimiteurs sur l'ensemble des itérations	Nombre de FN pour la $1^{ère}$ itération	Causes
Aucun	52%	14	7 Erreurs d'annotation
to	21%	11	2 Appositions
that	23%	12	2 Délimiteurs implicites
whether_if	3.3%	2	2 Ressources incomplètes
who_which	0.7%	0	4 Séquences reconnues uniquement par les automates
			1 Séquence ignorée par les règles et les automates
			2 Transducteur imprécis
			16 Descriptions non discriminantes
			3 Configurations spécifiques

TAB. 5.3 – Analyse des Faux Négatifs produits par le CB

Une règle de [Paice & Husk, 1987] reconnaît correctement la séquence, sa longueur de 7 mots entre le pronom et le délimiteur *who* fait décroître la probabilité mais la présence d'un adjectif, d'un verbe et d'un nom connus la rehausse au-delà du seuil d'acceptation. Le système classe donc ce pronom comme impersonnel alors qu'il est anaphorique. L'ajout de trois arcs entre le nœud `Longueur_Pronom_Delimiteur` et les nœuds

`Sequence_Contient_{nom,verbe,adjectif}` pourrait diminuer le poids de ces attributs relativement à la longueur des séquences et corriger ces erreurs.

La répartition des FP selon leurs délimiteurs sur l'ensemble des itérations nous apprend que les règles les moins fiables sont celles qui servent à la classification des séquences *it...to* : 47% des FP du système proviennent de ce type de règles. Une observation non exhaustive des erreurs commises par le système sur les autres itérations révèle un autre type d'erreur du système absente dans la première itération. Certaines séquences ont une configuration suffisamment spécifique, avec une faible longueur et la présence de mots connus comme *assumed* ou *shown*, pour être étiquetées *impersonnelles* par le système. Une décision qui est correcte lorsque le délimiteur est *that*, par exemple dans la séquence :

It is assumed that the SecY protein of B. subtilis has multiple roles...,

mais toujours erronée lorsque le délimiteur est *to* :

It is assumed to play a role in ...

Il faut étudier et traiter spécifiquement ces règles pour améliorer les performances de notre système.

Analyse détaillée des FN

Si le CB produit peu de FP, il produit en contrepartie plus de FN. Pour la 1^{re} itération, le CB produit 39 FN, le tableau 5.3 détaille les oublis du classifieur et présente les causes des oublis (nous donnons les phrases en annexe 1 dans le tableau 2).

7 erreurs du système sont dues à une erreur d'annotation dans le corpus annoté. Pour ces pronoms les valeurs réelles sont anaphoriques, les valeurs calculées par le système sont anaphoriques mais les valeurs dans le corpus annoté, qui sont fausses, sont impersonnelles. Il s'agit de pronoms anaphoriques qui ont été incorrectement annotés par un annotateur et dont la correction n'a pas été reportée dans le corpus ou encore de pronoms pour lesquels nous contestons à nouveau l'annotation. Nous laissons le lecteur libre d'apprécier nos choix présentés dans le tableau 2 de l'annexe 1 ou de les refuser.

2 erreurs du système sont dues à la présence d'une apposition. Ces pronoms sont impersonnels mais le système les étiquette anaphoriques. Prenons par exemple la séquence :

..., it is tempting –although still difficult– to draw conclusions about pathogenesis.

La présence de l'apposition *–although still difficult–* augmente la longueur de la séquence et diminue la probabilité pour le pronom d'être impersonnel. Seule une reconnaissance des appositions par les transducteurs que nous utilisons permetterait de corriger ces erreurs.

2 erreurs sont dues à un phénomène linguistique que nous n'avons pas représenté dans notre modèle : un délimiteur implicite. Considérons l'exemple :

Thus, it appears T3SO4 [*that*] *has no intrinsic...*

Le délimiteur *that* étant absent de la séquence aucune règle ne la reconnaît et le système classe le pronom anaphorique alors qu'il est impersonnel. Seule une inférence reposant sur une analyse linguistique plus fine aiderait à les reconnaître.

2 erreurs sont dues à l'incomplétude des ressources que le système utilise pour confirmer la classe du pronom *i.e.* les classes de verbes, d'adjectifs et d'adverbes. Dans la séquence :

It is remarkable that in B-CLL cells...

le pronom est impersonnel mais il est étiqueté anaphorique par le système. La présence de l'adjectif *remarkable* dans la classe des adjectifs connus aurait permis de changer la valeur de l'attribut `Sequence_Contient_adjectif` et de calculer une probabilité supérieure à 50%, ce qui aurait modifié la décision du système.

4 séquences sont reconnues uniquement par les automates et ignorées par les règles de Paice. Cette configuration obtient une probabilité trop faible pour être acceptée : le système classe donc le pronom anaphorique alors qu'il est impersonnel.

1 séquence a été ignorée à la fois par les automates et les règles en raison de la mise en apposition de l'adverbe *therefore* qui n'était pas prévue par nos transducteurs :

It is necessary, therefore, to understand the degree to which xenobiotics can disrupt endocrine systems.

Une étude linguistique des séquences ignorées par les règles de Paice devrait permettre de corriger ou de proposer de nouvelles règles capables de les reconnaître.

Nous retrouvons pour les dernières erreurs du système des causes identiques à celles de la section précédente :

2 pronoms impersonnels ont été étiquetés anaphoriques car nos transducteurs imprécis n'ont pas reconnu ces séquences,

16 pronoms impersonnels ont été décrits avec un vecteur non-discriminant dont la probabilité est inférieure à 50%,

3 pronoms impersonnels apparaissent dans une séquence de longueur supérieure à quatre mots, une configuration spécifique pour laquelle le système calcule une probabilité inférieure à 50%.

Conclusion

Lors de cette expérience nous avons montré que les résultats de notre CB etaient meilleurs que les résultats des systèmes concurrents. Nous avons ainsi établi, sur cette tâche, que l'exploitation d'un ensemble d'attributs pertinents bruités permet d'améliorer les performances du système. Cependant, cette expérience seule ne permet pas de comprendre le mécanisme de correction des attributs bruités par des attributs plus fiables, ni le mécanisme d'adaptation du système à un corpus d'un domaine différent. Pour comprendre ces deux mécanismes nous avons réalisé deux expériences supplémentaires que nous décrivons dans les sections ci-dessous.

5.3.2 La dépendance entre les attributs, un mécanisme pour les corriger

Afin de mettre en évidence le mécanisme de correction des attributs bruités par des attributs plus fiables, nous avons donc procédé à une deuxième expérience sur le corpus de génétique. Nous avons déterminé le rôle des dépendances dans la classification en analysant les différences entre les résultats de notre CB et ceux du CBN qui lui est associé. Par définition le classifieur bayésien naïf exploite les mêmes attributs que le classifieur bayésien mais aucun lien d'influence entre eux. A chaque itération de la validation croisée, nous avons calculé les fréquences nécessaires du corpus d'apprentissage pour paramétrer le CB et le CBN. Pour chaque occurrence du pronom *it* dans le corpus de test, nous avons considéré les décisions prises par le CB et le CBN. Nous avons calculé l'exactitude de chaque classifieur.

Malgré les bonnes performances du CBN, le CB obtient, à chaque itération, de meilleurs résultats. Sur la base de ces 20 couples de valeurs d'exactitude (la figure 5.4 chiffre les FP, les FN et l'Exactitude de chaque classifieur pour chaque itération), nous avons comparé leurs performances avec un test-t de Student-Fisher pour des données appariées. L'écart de performance observé en faveur du CB n'est pas dû au seul fait du hasard puisque la probabilité d'un tel écart est de $3,1 * 10^{-16}$ avec un seuil de signification de 0.001.

Privé des relations de dépendance entre les attributs, le CBN surestime leurs fiabilités. L'hypothèse d'indépendance conduit le CBN à calculer des probabilités trop élevées et à produire un nombre important de FP. Ces FP proviennent principalement des occurrences reconnues uniquement par une règle de [Paice & Husk, 1987] de la forme *it...to*. Lors de la 1^{re} itération, sur les 91 FP du CBN, on compte 70 cas de ce type. Sur la simple observation qu'une règle [Paice & Husk, 1987] de la forme *it...to* reconnaît une séquence, le CBN estime que le pronom contenu dans la séquence est impersonnel à 98.8%, alors que le CB, sur ces mêmes observations, donne une probabilité de 45.4% seulement. Les multiples liens d'influence dont dépend le nœud `Regle_Paice` expriment un ensemble plus riche de contraintes, une information dont le CB tient compte. Il donne des probabilités plus faibles et ne fait que 10 erreurs de ce type. Le CB n'accepte ces séquences que si elles contiennent un verbe et un adjectif connus et si le pronom et le délimiteur sont à moins de 5 mots de distance. Même si elles produisent des erreurs, ces descriptions se trouvent, dans notre représentation, justifiées, car elles décrivent aussi 3 occurrences impersonnelles, par exemple l'occurrence de la séquence

However, it often appears difficult to establish cell lines or transgenic animals....

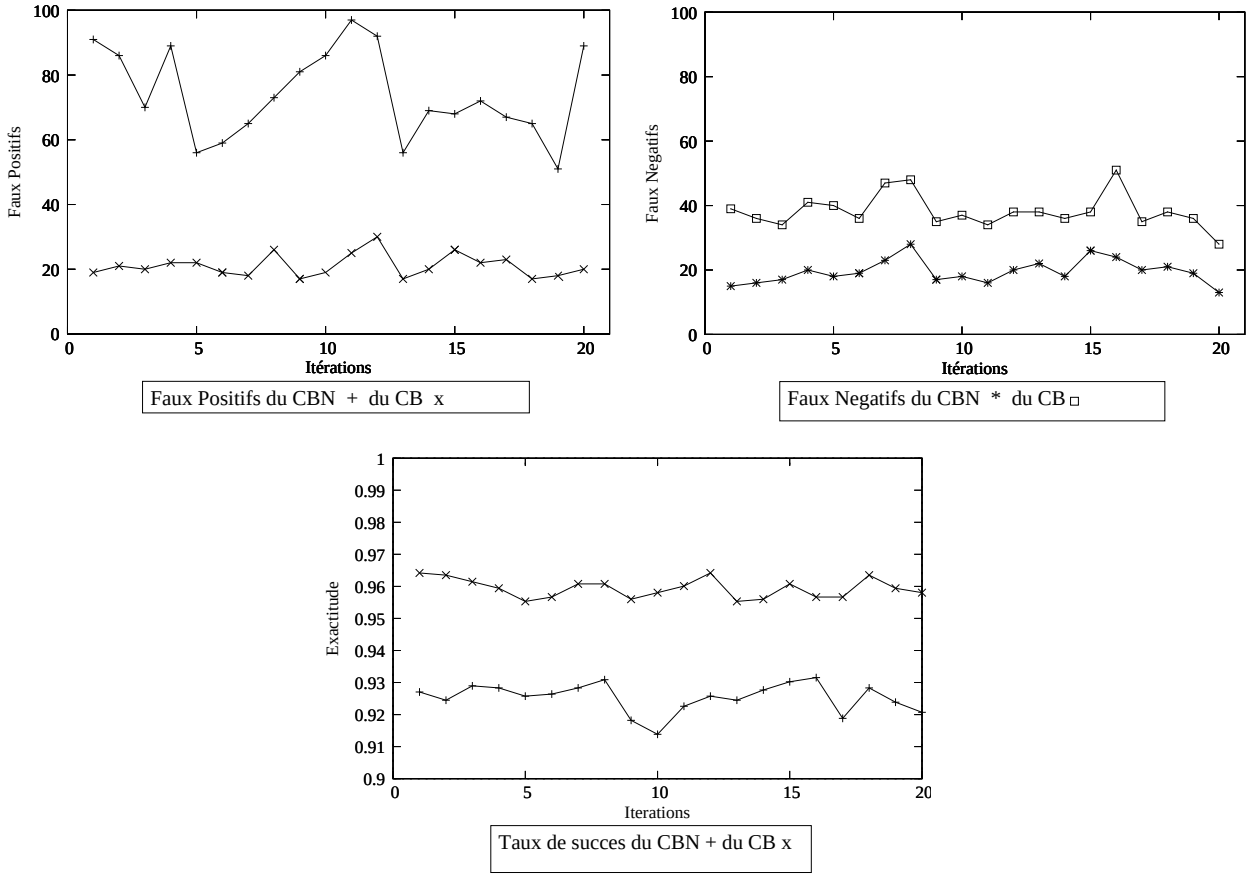

FIG. 5.4 – Faux Positifs/Négatifs et Exactitude des classifieurs de pronoms impersonnels CB et CBN pour chaque itération

	Exactitude	Faux Positifs	Faux Négatifs
Paramétrage Génomique	95.9	10	12
Paramétrage Management	94.8	15	13

TAB. 5.4 – Résultats du classifieur bayésien sur le corpus Génomique avec différents paramétrages

Phrases	Délimiteur	Analyse
Currently *it* is accepted the existence of a cytoplasmic and/or nuclear receptor, without explaining satisfactorily *how* the hormones come to the nucleus.	how	
It was found, in particular, *that* the element TCEd, which binds the transcription factor NF-kB, is very active in all three B clones tested	that	Poids insuffisant des règles de [Paice & Husk, 1987] reconnaissant des séquences longues
It is, however, still unclear *whether* LTB4 acts in this regard directly or indirectly by stimulating the release of chemotactic and inflammatory cytokines.	wether	

TAB. 5.5 – Faux Négatifs produits par le CB avec le paramétrage Génomique et correctement classés par le CB avec le paramétrage Management

5.3.3 Le paramétrage des attributs, un mécanisme d'adaptation au corpus

Les liens d'influence entre les variables expriment un ensemble de contraintes attendues par le classifieur pour calculer la classe la plus probable, mais ces liens ne sont pas suffisants. La mesure de la fiabilité *a priori* des attributs est nécessaire pour exprimer la nature de la contrainte.

Or cette fiabilité *a priori* varie d'un corpus à l'autre, en fonction de leur genre et de leur domaine. Nous avons mené une nouvelle expérience pour déterminer la robustesse de notre classifieur au changement de corpus. Pour cette dernière expérience nous avons modifié le protocole d'expérimentation. Dans un premier temps nous avons paramétré le CB à partir des fréquences calculées sur le corpus Management puis nous avons appliqué le CB, ainsi paramétré, sur 548 occurrences tirées aléatoirement du corpus de Génétique. Dans un deuxième temps nous avons reparamétré le CB sur 671 occurrences du corpus de Génétique, disjointes des 548 occurrences précédentes. Avec ce nouveau paramétrage, nous avons appliqué le CB sur les 548 occurrences précédentes du corpus de Génétique.

Le tableau 5.4 résume les résultats obtenus par le CB selon les différents paramétrages. L'écart des performances du CB confirme que la fiabilité des attributs diffère en fonction du domaine du corpus et que l'étape de paramétrage offre un mécanisme élégant pour adapter le classifieur à de nouveaux corpus. En recalculant les fréquences sur un nouveau corpus d'entraînement, nous réévaluons la fiabilité de chaque attribut. Possédant une mesure précise de la

Phrases	Délimiteur	Analyse
It therefore was of interest to establish which, *if* any, commitment events were affected by oxidative signalling during cell cycle entry.	if	
It now remains to be investigated *which* parameters determine the repertoire of the memory response and possibly restrict its diversity after subsequent antigenic challenges.	which	Poids des mots connus insuffisant
Thus, *it* is not unexpected that this versatile cellular homeostatic switch would be affected by a variety of viral pathogens, *which* have evolved mechanisms to utilize various aspects of Rel/NF-kappa B activity to facilitate their replication, cell survival and possibly evasion of immune responses.	which	

TAB. 5.6 – Faux Négatifs produits par le CB avec le paramétrage Management et correctement classés par le CB avec le paramétrage Génomique

Phrases	Délimiteur	Analyse
Overexpression of SRF in B cells causes the IL-2R enhancer to function as well as *it* does in T cells, suggesting *that* the high level of SRF binding in T cells is functionally important.	that	
Although this reversal by IL-2 was pronounced, *it* was not complete, suggesting *that* IL-10 may have some effects not directly related to the suppression of IL-2 production.	that	
However, *it* was still lower than that in healthy subjects (P<0.01).	that	Poids des règles de [Paice & Husk, 1987] reconnaissant des séquences longues trop important
The sequence between position -42 and -76 base pairs (bp) was required for efficient transcription in cells that express alpha 4, but *it* showed no activity in HeLa cells, which do not express alpha 4.	which-who	
We have found that *it* is phosphorylated, predominantly on serine residues, *when* transiently overexpressed in 293 cells.	wether-if	

TAB. 5.7 – Faux Positifs produit par le CB avec le paramétrage Management et correctement classés par le CB avec le paramétrage Génomique

perte ou du gain de fiabilité pour chaque attribut, nous pouvons privilégier les attributs les plus fiables lors de l'étape d'inférence.

Dans cette expérience le système est identique, seul le paramétrage varie. Le système produit des erreurs que le changement de paramétrage ne corrige pas, les tableaux 5.5, 5.6 et 5.7 ne les mentionnent pas. Ces erreurs relèvent d'une autre cause que celle d'une mauvaise évaluation de la fiabilité des attributs (que le système soit paramétré par un corpus ou l'autre, il classe toujours une séquence où le délimiteur est implicite comme anaphorique). Nous nous focalisons sur les erreurs qui dépendent du paramétrage et qui sont détaillées dans les précédents tableaux.

Nous expliquons les FN du système paramétré avec le corpus génétique par une cause unique : le poids d'une règle de [Paice & Husk, 1987] est sous-estimée dans notre paramétrage lorsque la séquence finit par un délimiteur de la classe *whether-if*[42] et a une longueur comprise entre 4 et 6 mots. Sur ces seules données observées le réseau bayésien évalue la probabilité du pronom d'être impersonnel entre 8.8% et 19% alors qu'avec le paramétrage de Management les probabilités sont nettement supérieures, entre 38% et 99%.

Cet écart entre les probabilités précédentes provient essentiellement d'une variation linguistique entre les deux corpus. En effet, pour le corpus Management, 34.6% des séquences introduisant un pronom impersonnel ont une longueur supérieure ou égale à 4 mots contre seulement 10.0% pour le corpus Génomique[43]. La conséquence de cette variation linguistique est une sur-évaluation du poids des valeurs de l'attribut `Longueur_Pronom_Delimiteur` par le système paramétré sur le corpus Management, ce qui entraîne l'étiquetage des 5 faux positifs du tableau 5.7. Le système paramétré avec la partie du corpus Génomique est plus prudent avec les séquences de longueur supérieure à trois mots et attend la présence d'un adjectif, d'un nom ou d'un verbe pour les accepter[44].

Les 3 FN du système paramétré sur le corpus Management sont dûs à une erreur d'implémentation qui a déjà été mentionnée dans la section 5.3.1 : le système ne s'arrête pas au premier délimiteur rencontré mais au suivant. La correction de cette erreur devrait aussi corriger les erreurs du système. Toutefois, il est intéressant de noter que le système avec les paramètres adaptés classe correctement ces séquences. La raison est simple, elles contiennent toutes un adjectif, un verbe ou un nom connus ce qui réhausse suffisamment la probabilité pour le pronom d'être impersonnel. Les verbes et les adjectifs sont aussi rencontrés dans le corpus Management mais moins souvent que dans le corpus Génomique, leur poids est donc légèrement plus faible.
Dans le corpus Génomique :
P(Pronom=Impersonnel|Sequence_Contient_Verbe=Connu)=0.954
P(Pronom=Impersonnel|Sequence_Contient_Adejctif=Connu)=0.784
Dans le corpus Management :
P(Pronom=Impersonnel|Sequence_Contient_Verbe=Connu)=0.923

[42] Cette classe comprend les délimiteurs : *whether, why, if, where, when, what, how*.

[43] Un écart très sensible sur les séquences comprenant quatre mots entre le pronom impersonnel et le délimiteur, le corpus Management compte 22 séquences contre seulement 7 sur le corpus Génomique.

[44] Pour chiffrer le phénomène comparons les 2 probabilités obtenues sur Management :
- *P(Pronom=impersonnel|Longueur_Pronom_Delimiteur=quatre)=0.60*
- *P(Pronom=impersonnel|Longueur_Pronom_Delimiteur=six)=0.10*

avec les 2 probabilités obtenues sur Génomique :
- *P(Pronom=impersonnel|Longueur_Pronom_Delimiteur=quatre)=0.22*
- *P(Pronom=impersonnel|Longueur_Pronom_Delimiteur=six)=0.01*

P(Pronom=Impersonnel|Sequence_Contient_Adejctif=Connu)=0.755[45].

L'analyse des erreurs du système selon ces différents paramétrages nous a conduit à interroger les probabilités de notre réseau bayésien. Les réponses obtenues nous ont permis de caractériser quantitativement deux variations linguistiques entre nos corpus : une classe de verbes et d'adjectifs légèrement différents et une longueur des séquences impersonnelles plus importante dans le corpus Management. En poursuivant ces investigations, nous pourrions en caractériser bien plus. Ces connaissances, après avoir été soumises à une étude linguistique, pourraient nous servir à prévenir l'impact d'une connaissance selon le domaine du corpus. Prenons par exemple les séquences terminées par le délimiteur *to*. Leur faible fiabilité nous a en premier lieu poussé à les suspecter dans la baisse observée des performances lors de cette expérience. Ces séquences reconnaissent moins d'occurrences et sont beaucoup moins fiables sur le corpus Génomique que sur le corpus Management. Parmi les séquences reconnues par une règle de [Paice & Husk, 1987] dans le corpus Génétique, 25%[46] le sont par une règle terminée par le délimiteur *to* contre 45% dans le corpus Management. Sur ce dernier corpus, ces règles produisent seulement 40%[47] de faux positifs contre 59% sur le corpus Génomique. Si la présence de l'attribut augmente la probabilité pour le pronom d'être impersonnel sur le corpus Management, elle la diminue sur le corpus Génomique, en l'absence d'observation complémentaire.

De telles connaissances accentuent encore la nécessité d'une phase de paramétrage du système et montrent indirectement l'efficacité du mécanisme de renforcement de notre classifieur. Le poids de la valeur *To* de l'attribut calculé sur le corpus Management est sur-évalué pour le corpus Génomique et aurait dû provoquer de nombreuses erreurs, mais le système a corrigé cette mauvaise information grâce aux autres attributs.

5.4 Conclusion

L'objectif de ce chapitre était de valider notre approche sur un problème plus simple que celui du choix de l'antécédent, la distinction des pronoms impersonnels des pronoms anaphoriques. Ce problème est parfaitement adapté. D'une part parce que cette distinction est une étape nécessaire pour la résolution des anaphores pronominales. Le système de distinction a été incorporé au système de résolution des anaphores dans le chapitre 6. D'autre part, parce que nous retrouvons l'opposition des connaissances linguistiques complexes et des indices de surface, bien que les attributs nécessaires à la distinction soient peu nombreux et relativement fiables (même pour les attributs traduisant les connaissances complexes).

L'état de l'art distingue deux stratégies qui selon nous sont complémentaires. Les deux premiers systèmes que nous avons décrits reposent sur des règles syntaxiques pour discriminer les séquences introduisant les pronoms impersonnels. Les règles du premier système, le système de [Lappin & Leass, 1994], sont très contraignantes et ont un mauvais rappel, alors que les règles du second système, le système de [Paice & Husk, 1987], qui ne décrivent plus aussi précisement les séquences, sont moins précises. Les deux derniers systèmes, les systèmes de [Evans, 2001] et de [Litran *et al.*, 2004], s'appuient sur des indices de surface uniquement et un apprentissage

[45] Curieusement le poids d'un nom connu est plus fort sur le corpus Management que sur le corpus Génomique : *P(Pronom=Impersonnel|Sequence_Contient_Nom=Connu)=0.553* contre 0.524 respectivement.

[46] *i.e. P(Delimiteur=To|Regle_Paice=Reconnu)=0.25*

[47] *i.e.P(Delimiteur=To et Pronom=Anaphorique|Regle_Paice=Reconnu)=0.45*

des poids de ces indices pour discriminer les pronoms impersonnels.

Après avoir formalisé le problème de la distinction des pronoms comme un problème de classification, nous avons proposé un classifieur bayésien reposant sur le modèle des réseaux bayésiens. Ce modèle nous permet d'intégrer les règles syntaxiques et les indices de surface dans une même représentation. Nous mesurons avec précision la fiabilité de chaque attribut sur le corpus grâce au paramétrage du modèle. Nous tirons profit du mécanisme d'inférence pour corriger les défauts propres de chaque approche grâce aux qualités des autres. Le système accepte avec une forte probabilité les pronoms impersonnels reconnus par une règle de [Lappin & Leass, 1994]. Mais ces règles reconnaissent peu de pronoms : en complétant ces règles par les règles de [Paice & Husk, 1987], nous corrigeons ce problème de rappel. Les règles de [Paice & Husk, 1987] reconnaissent la plupart des pronoms impersonnels mais, malheureusement, aussi des pronoms anaphoriques. Pour distinguer les pronoms impersonnels et anaphoriques nous corroborons la décision des règles de [Paice & Husk, 1987] grâce à la présence des indices de surface des systèmes [Evans, 2001] et de [Litran *et al.*, 2004].

Nous avons évalué avec succès notre système. Le classifieur fondé sur un réseau bayésien réalise une meilleure classification des pronoms impersonnels que les systèmes de [Lappin & Leass, 1994], [Paice & Husk, 1987] et [Litran *et al.*, 2004] avec lesquels nous l'avons comparé.

Durant cette expérience nous avons aussi mis en évidence l'intérêt des relations de dépendances entre les attributs du réseau. Le classifieur bayésien obtient de meilleurs résultats que le classifieur bayésien naïf qui lui est associé. Les liens de dépendance qui, nous le rappelons, sont des probabilités conditionnelles, expriment des contraintes que le mécanisme d'inférence emploie pour corroborer la fiabilité d'un attribut par l'observation des valeurs des attributs dont il dépend.

Enfin, en substituant au corpus d'apprentissage un corpus d'un autre domaine, nous avons montré l'intérêt de paramétrer le modèle au domaine du corpus d'expérimentation. Les phénomènes linguistiques varient d'un domaine à l'autre et peuvent modifier la fiabilité des attributs du système. Le rôle de l'étape de paramétrage est de modifier les probabilités conditionnelles du modèle. Cette étape nous assure, dans une certaine mesure, que les contraintes du modèle expriment les nouveaux phénomènes linguistiques et qu'elle adapte le poids des attributs au corpus, garantissant derechef le bon fonctionnement du mécanisme d'inférence.

Chapitre 6

Un classifieur bayésien pour la résolution des anaphores

Sommaire

Ayant vérifié le bon comportement de notre modèle sur un problème "simple" nous pouvons généraliser notre approche au problème complet de la résolution des anaphores. Cette section décrit notre classifieur et rapporte les performances de ce dernier sur deux corpus de natures différentes.

6.1 Une approche intégrée pour la résolution des pronoms anaphoriques

Si les indices approchés proposés lors des années 1990 ont permis l'implémentation de systèmes robustes [Mitkov, 2002], les travaux récents commencent à en mesurer les limites. L'étude de [Kehler *et al.*, 2004] montre ainsi que les patrons de collocations de [Dagan & Itai, 1990] n'améliorent pas les performances d'un système qui exploite déjà des informations

morpho-syntaxiques. Les auteurs en concluent que l'apport des fréquences tient davantage du hasard que d'une véritable capture du sens sémantique.

Les limites rencontrées par les systèmes à base d'indices de surface nous renvoient au problème initial. Nous avons besoin de connaissances sémantiques et syntaxiques complexes pour la résolution de l'anaphore pronominale. Ces connaissances linguistiques, lorsqu'elles sont disponibles, ne sont pas fiables. On peut chercher à les remplacer par des indices de surface dont le calcul est toujours réalisable et plus fiable mais ces indices peuvent ne pas exprimer, ou seulement de manière imprécise, les connaissances nécessaires à la résolution, ce qui produit des erreurs.

Nous proposons à nouveau un modèle reposant sur les Réseaux Bayésiens. Ce modèle offre la possibilité d'unifier dans une unique représentation connaissances linguistiques et indices de surface. Cette unification devrait permettre de corroborer les connaissances linguistiques grâce aux indices de surface qui sont observés en corpus. A l'inverse, l'exploitation de connaissances linguistiques pourrait corriger certaines erreurs des systèmes à base d'indices de surface.

6.1.1 La résolution des anaphores vue comme un problème de classification

Le choix de l'antécédent, tout comme la distinction des pronoms impersonnels, se reformule en un problème de classification :

Soit,

- *Corpus* un ensemble de textes d'un même domaine,
- *Corpus_entrainement* et *Corpus_test* deux sous-ensembles stricts disjoints de *Corpus*,
- $Pronouns$ et $Candidats$, les ensembles des occurrences des pronoms et des SN candidats présents dans $Corpus$

On définit $Antecedent$ et $Not\_Antecedent$ les classes complémentaires des couples e_i formés par un pronom et un candidat de l'ensemble $Pronoun \times Candidat$; e_i appartient à la classe $Antecedent$ si le candidat est l'antécédent du pronom et à la classe $Not\_Antecedent$ si le candidat n'est pas l'antécédent ou si le pronom est impersonnel. Tout couple e_i est décrit par un vecteur $v_i = a_1, ..., a_v$ d'attributs. Chaque attribut a_i est sélectionné sur la base d'une analyse humaine de *Corpus_entrainement* et représente une connaissance linguistique ou un indice de surface.

Considérons la phrase :

No doubt you made [a contribution]$_1$ and I haven't mentioned [it]$_1$.

elle contient 4 couples : $e_1 = (it_1, doubt), e_2 = (it_1, you), e_3 = (it_1,$ ***a contribution***$), e_4 = (it_1, I)$

parmis lesquels seul $e_3 = (it_1,$ ***a contribution***$)$ appartient à la classe $Antecedent$, les autres couples appartiennent à la classe complémentaire. Dans cet exemple nous décrivons chaque couple au moyen d'un vecteur de 3 attributs : le rang du candidat dans la liste des candidats de la phrase, son rôle syntaxique et le rôle syntaxique du pronom. Le vecteur décrivant le couple e_3 est donc $v_3 = (3,$ ***complément, complément***$)$ car ***a contribution*** est le troisième candidat de la phrase et il est tout comme le pronom complément d'un verbe.

Pour toute occurrence d'un pronom p de *Corpus_test*, un classifieur bayésien attribue une probabilité d'appartenir à la classe $Antecedent$ à chaque couple présent dans la fenêtre de re-

cherche où le pronom p apparaît. Si le pronom est anaphorique, le candidat du couple qui obtient la plus grande probabilité est désigné comme antécédent. Dans le cas d'une égalité entre plusieurs couples nous appliquons les heuristiques de [Mitkov, 2002][48] pour choisir l'antécédent.

6.1.2 Modélisation et emploi du classifieur bayésien

Pour réaliser le classifieur de la figure 6.1 nous avons procédé en deux temps. Nous avons commencé par implémenter la première version du système MARS (voir 2.2.3) dont nous avons conservé tous les indices de surface à l'exception des indices *référence immédiate* et *Instruction séquentielle* qui sont spécifiques au domaine des textes techniques pour lesquels il a été développé. Cette réimplémentation du système nous servira de système de comparaison lors de l'évaluation de notre classifieur dans la section 6.3.

Nous avons ensuite implémenté notre classifieur. Nous utilisons les indices de notre implémentation du système MARS auxquels nous avons ajouté une série d'indices supplémentaires, pertinents pour le calcul de la saillance, exploités par d'autres travaux de l'état de l'art. Nous avons complété les indices de surface du système MARS, lorsque cela était possible, par les connaissances linguistiques qu'ils renforcent. Par exemple, le sujet d'une phrase est souvent l'élément saillant mais comme le calcul du rôle grammatical peut être erroné, il est intéressant d'exploiter en parallèle l'information concernant un indice de surface : le fait pour un groupe nominal d'être le premier de la phrase peut confirmer ou infirmer l'hypothèse du rôle grammatical, les sujets tendant en effet à figurer en début de phrase. Les attributs décrivant les indices de notre système MARS sont colorés en noir. Les attributs qui les complètent sont en gris. Le noeud de prédiction est le noeud *Candidate*, au centre. Il estime la probabilité pour une occurrence d'un candidat d'être l'antécédent d'un pronom donné.

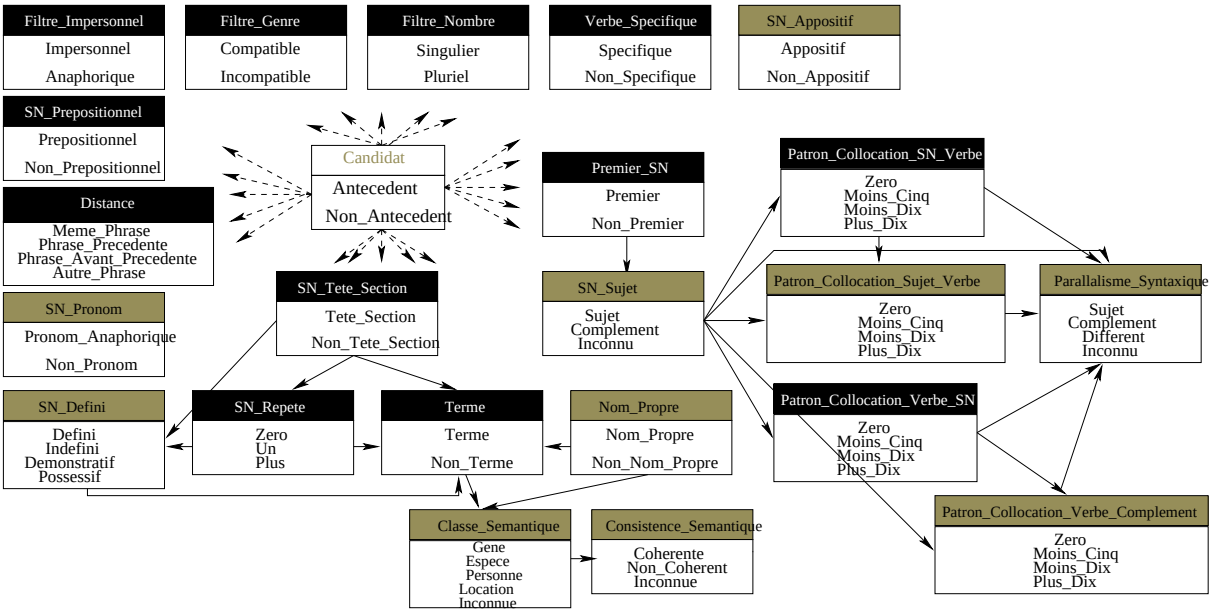

FIG. 6.1 – Un réseau bayésien pour la classification des candidats antécédents

[48] voir la section 2.2.3.

Nous conservons les contraintes du système MARS *i.e.* les cohérences de genre, de nombre, le filtre des pronoms impersonnels. La liste suivante décrit les attributs qui sont pris en compte dans notre classifieur et explique comment ils sont calculés :

- `Filtre_Genre` : l'évaluation du système ayant été effectuée sur des textes techniques et de génomique (voir la section 6.1.3) où le genre neutre domine, nous avons réduit le filtre du genre à sa plus simple expression, il élimine uniquement les pronoms incompatibles : *we, they, you,I, she, he*.
- `Filtre_Nombre` : le filtre élimine les pronoms pluriels et les SN composés d'un unique mot au pluriel. Pour les SN composés, l'heuristique de [Cardie & Wagstaff, 1999] s'étant révélée très bruitée sur notre corpus de développement, nous lui en avons préféré une autre : tout SN composé qui contient un nom pluriel est rejeté.
- `Filtre_Pronom_Impersonnel` : pour distinguer les pronoms impersonnels nous avons appliqué le classifieur décrit dans le chapitre précédent. Pour faciliter la programmation, les deux réseaux n'ont pas été unifiés mais ils auraient pu l'être.
- `Premier_SN` : la valeur de cet attribut dépend de la position du pronom. Nous prenons le premier SN de la phrase précédente, lorsque le pronom est en début de phrase et le premier SN de la phrase courante lorsque le pronom ne débute pas la phrase.
- `SN_Sujet` : chaque candidat et chaque pronom doivent avoir un rôle grammatical associé mais ce calcul est encore très silencieux. Nous avons ajouté la valeur *Inconnu* afin de différer l'intégration des rôles inconnus dans l'évaluation de notre système.
  Comme nous l'avons vu ci-dessus cet attribut est renforcé (ou diminué) par l'attribut `Premier_SN`.
- `Verbe_Specifique` : nous avons reproduit la classe des verbes spécifiques publiés dans [Mitkov, 1998] que nous avons complétée manuellement par une liste de verbes du domaine de la génomique extraits d'un corpus d'entraînement.
- `SN_Repete` : nous repérons les répétitions uniquement par des comparaisons entre les chaînes des caractères des constituants et entre les têtes des constituants de notre analyse. Pour calculer la tête d'un candidat nous appliquons une heuristique simple : si le candidat est précédé d'un article, la tête sera composée des premiers noms successifs, sinon par les derniers noms successifs. Par exemple, la tête du candidat *The overall properties of Mep1p* sera *properties*.
  Ce nœud est lié au nœud *Candidat_Tete_Section* car les candidats présents dans les titres de sections ont de très fortes probabilités d'être répétés dans le corps des sections.
- `SN_Tete_Section` : nous marquons les candidats qui apparaissent dans les titres de sections. Les sections ne sont pas calculées, elles sont données par la structure du document.
- `Patron_Collocation_SN_Verbe` et `Patron_Collocation_Verbe_SN` : les patrons de collocations sont de la forme <*SN/pronom verbe*> et <*verbe SN/pronom*> où SN est le premier syntagme nominal (resp. le premier pronom *it*) qui précède le verbe pour la première règle et qui le suit pour la deuxième règle. Nous considérons la forme lemmatisée du verbe dans les patrons. Nous calculons les fréquences sur la régularité des têtes des candidats et sur l'ensemble du corpus d'entraînement. Considérons le contexte de l'anaphore suivante :
  [Another ORF]$_1$, apparently in the same transcription unit, was found downstream from the amylase gene. [It]$_1$ encoded a protein that was closely related to the maltose-binding

protein of Escherichia coli.

Pour la résoudre, le système cherche successivement les fréquences des patrons de collocations où apparaissent les têtes des 3 candidats possibles : <*ORF/encode*>, <*amylase gene/encode*> et <*transcription unit/encode*>. Seule la collocation <*ORF/encode*> est retrouvée dans différentes phrases du corpus[49].

– `Patron_Collocation_Sujet_Verbe` et `Patron_Collocation_Verbe_Complement` : Nous complétons les patrons de collocations précédents par de nouveaux patrons de collocations de la forme <*Sujet-verbe*> et <*verbe-complement*>. Ces patrons sont plus fiables que les patrons précédents : le candidat ne partage une collocation identique au pronom que s'il a déjà été le sujet (resp. le complément) du verbe dont le pronom est lui-même le sujet (resp. le complément). Si nous reprenons le contexte précédent de l'anaphore :

We show that in B. subtilis the ORF preceding the rpsA homologue encodes a protein which is highly similar to the product of the E. coli mssA gene which is located upstream of rpsA.

la collocation <*ORF/encode*> de la phrase n'est pas retrouvée par le patron de collocation <*SN/verbe*> car le verbe *preceding* suit immédiatement le candidat *ORF* : la collocation retrouvée est donc <*ORF/precede*>. Si le système connaît le rôle grammatical du candidat et du pronom, il retrouve la collocation <*ORF-encode*> avec le patron <*Sujet-verbe*>.

– `Parallelisme_Syntaxique` : nous comparons les rôles syntaxiques du pronom et du candidat lorsque ce dernier est connu, sinon la valeur est posée à *Inconnu*.

Nous avons lié le nœud *SN_Sujet* et les nœuds des patrons de collocations avec l'objectif de renforcer (ou de diminuer) le poids du parallélisme syntaxique entre le pronom et le candidat. Supposons que le pronom précède immédiatement le verbe dont il est le sujet et que le candidat soit le sujet de sa proposition. Le système va rechercher les collocations <*NP/verbe*> ou <*sujet-verbe*>. S'il trouve plusieurs collocations dans le corpus où le candidat est le sujet du verbe suivant le pronom, la probabilité pour le candidat d'être parallèle syntaxiquement doit être augmentée ainsi que la fiabilité du rôle grammatical du candidat.

– `Terme` : nous avons calculé les termes du domaine d'une manière différente pour chaque expérience que nous avons menées. Pour l'expérience sur un corpus de génomique nous avons appliqué une ressource terminologique du domaine alors que pour l'expérience sur un corpus technique nous avons recalculé les termes sur le corpus.

La propriété pour un syntagme d'être un terme du domaine n'est pas une propriété précise. Un syntagme peut être plus ou moins représentatif du domaine. Pour cette raison, nous avons choisi de renforcer la fiabilité pour un candidat d'être un terme par différents critères d'importance discursive du candidat. Ainsi le fait que le candidat apparaisse dans un titre de section, qu'il soit démonstratif ou encore qu'il soit répété dans le document contribue à augmenter sa probabilité d'être un terme. Nous considérons aussi que les noms propres du document sont des termes du domaine, comme par exemple le nom

[49] Par exemple :

Complementation studies using an E. coli bioF mutant and a B. subtilis bio112TG3 strain, revealed that the third ORF of this cluster encodes 7-keto-8-aminopelargonic acid synthetase.

This ORF encoded a predicted 461-residue product which had high identity with Class I sugar transporters of the major facilitator superfamily.

d'espèce *Bacillus subtilis* pour le domaine de la génomique ou le nom d'un logiciel pour le domaine informatique.

- `SN_Defini` : nous supposons qu'un NP est indéfini s'il n'est pas précédé par un article défini, possessif ou démonstratif.
  Ce nœud est lié aux nœuds *SN_Defini* et *SN_Tete_Section* car un candidat qui apparaît dans un titre de section ou qui est répété dans le document est un candidat qui a déjà été introduit dans le discours et, doit, normalement, être défini.
- `SN_Prepositionnel` : nous calculons les prépositions à partir de l'analyse en constituants du texte, un syntagme est considéré comme prépositionnel s'il suit immédiatement une préposition. Dans l'analyse syntaxique suivante,
  (S (NP (NP Point mutations)(PP in (NP the CRE sequence)))(VP caused (NP (NP the loss) (PP of (NP (NP catabolite repression) (PP of (NP the operon))))))
  les syntagmes *the CRE sequence, the operon, catabolite repression* et *catabolite repression of the operon* sont prépositionnels.
- `Distance` : en raison des analyses syntaxiques partielles des phrases complexes de notre outil nous avons décidé de simplifier le calcul de distance entre le pronom et le candidat au niveau de la phrase et non de la proposition. Dans le contexte suivant :
  We demonstrate that a ygaG mutant has the perR phenotype. It is highly resistant to peroxides and overexpresses catalase, alkyl hydroperoxide reductase and the DNA binding protein MrgA.
  le candidat *a ygaG mutant* est dans la proposition précédant le pronom et le candidat *We* dans celle d'avant alors qu'ils sont tous les deux dans la même phrase.
- `Nom_Propre` : nous utilisons l'analyseur morpho-syntaxique de notre système pour reconnaître les noms propres.
- `SN_Pronom` : pour retrouver les pronoms *it* nous nous fions aussi aux étiquettes de l'analyseur morpho-syntaxique. Le système garde en mémoire le statut anaphorique ou impersonnel des pronoms qu'il a déjà traités. Il peut donc calculer la valeur pronom anaphorique de l'attribut lorsque le candidat d'un pronom anaphorique est aussi un pronom anaphorique (les pronoms qui ont été jugés impersonnels par le système sont éliminés de la liste des candidats) comme dans l'exemple suivant :
  Bacillus subtilis is unable to grow by consuming galactose because [it]$_{candidat}$ is unable to transport [it]$_{anaphore}$ into the cell.
- `SN_Appositif` : nous trouvons les appositions par une heuristique simple. Le candidat ne doit pas contenir de verbe et être entouré d'un signe de ponctuation double différent des parenthèses.
- `Classe_Semantique` : nous nous appuyons sur un outil de reconnaissance des entités nommées qui associe la classe sémantique à l'entité nommée annotée. Notre outil étant spécialisé pour le domaine de la génomique pour les deux expériences, il recherche peu de classes sémantiques.
  La fiabilité d'une certaine classe sémantique associée à un candidat peut être renforcée si le candidat est un nom propre ou encore un terme. Les classes sémantiques *espèce* ou *personne* sont par exemple corroborées lorsque l'analyseur morpho-syntaxique étiquette le candidat comme nom propre. De même certains gènes (ex. *open reading frame (ORF)*) ou noms d'espèces reconnus comme termes du domaine renforcent la classe sémantique associée au candidat.

– `Coherence_Semantique` : pour vérifier la cohérence sémantique d'un candidat nous généralisons les patrons de collocations du type <*sujet-verbe*>. Nous recherchons l'ensemble des classes sémantiques des sujets retrouvés dans un patron de collocations identique à celui du pronom, si la classe du candidat appartient à cet ensemble, les classes sémantiques du pronom et du candidat sont cohérentes. Considérons à nouveau le contexte de l'anaphore ci-dessus. Dans cet exemple nous connaissons la classe sémantique du candidat *Another ORF* : c'est un gène. Avec cette information plus générale nous recherchons le nombre de collocations où un nom de gène est sujet du verbe dont le pronom est lui-même sujet. Ainsi le patron <*GENE/encode*> sélectionne des phrases comme :
The $[sacT]_{gene}$ *gene which controls the sacPA operon of Bacillus subtilis [encodes] a polypeptide homologous to the B. subtilis SacY and the Escherichia coli BglG antiterminators.*
$[sacX]_{gene}$ *and* $[sacY]_{gene}$ *[encode] respectively a negative and a positive regulator involved in induction by sucrose of the exoenzyme levansucrase.*
Malheureusement nous ne pouvons calculer la cohérence sémantique d'un candidat et du pronom que si le candidat est une entité nommée reconnue. La fiabilité de la cohérence sémantique est donc conditionnée par celle de la classe sémantique du candidat.
– `Candidat` : nous avons lié le nœud de prédiction avec tous les nœuds du réseau car chaque nœud intervient dans le calcul de la saillance du candidat.

Comme dans le chapitre précédent pour l'identification des pronoms impersonnels, la structure du réseau a été imposée par un expert et n'a pas été apprise automatiquement. Les probabilités ont été apprises avec l'approche du maximum de vraisemblance (voir la section 4.1.3) sur un corpus d'apprentissage. Les valeurs de certaines informations sont inévitablement bruitées. Par exemple certains antécédents sont supposés être des SN pluriels en raison d'un mauvais étiquetage morpho-syntaxique ou d'un mauvais calcul du nombre. La valeur *a priori* de la probabilité conditionnelle *P(Filtre_Nombre = Incompatible|Candidat = Antecedent) =0.02* est donc non nulle[50] et, malgré cette faible probabilité, elle permet au système, durant la phase de test, d'accepter certains candidats qui auraient du être rejetés par le filtre. Après une étude de ces probabilités bruitées, les experts pourraient corriger les données et les intégrer aux données d'apprentissage avec l'approche du *Maximum a posteriori* afin d'atténuer le bruit des paramètres.

6.1.3 Corpus et ressources

Les corpus d'expériences

Corpus MARS : Dans le but de comparer les performances de notre réimplémentation du système MARS nous avons appliqué notre système sur un corpus de manuels techniques qui était distribué par les auteurs sur le site d'évaluation du système MARS. Le corpus n'est maintenant plus disponible sur le site. Le corpus distribué présente des différences avec le corpus d'évaluation de [Mitkov, 1998] qui comprend un document supplémentaire. Il y a aussi 4 documents différents entre les deux corpus bien que les documents soient toujours des manuels techniques. Le tableau 6.1 chiffre les caractéristiques du corpus. Nous avons nettoyé les annotations coréférentielles inutiles pour notre tâche et recodé les pronoms impersonnels et les re-

[50] Le même phénomène est aussi observé pour le filtre du genre.

Nombre de manuels	Mots	Nombre de pronoms	Pronoms Anaphoriques	Pronoms Impersonnels	Anaphores Interphrastiques	Anaphores Intraphrastiques
7	39 500	288	54	234	144	90

TAB. 6.1 – Caractéristiques du corpus MARS

Nombre de résumés	Mots	Nombre de pronoms	Pronoms Anaphoriques	Pronoms Impersonnels	Anaphores Interphrastiques	Anaphores Intraphrastiques
2209	800 000	671	429	242	164	265

TAB. 6.2 – Caractéristiques du corpus Transcript

lations anaphoriques des pronoms *it* restants avec un schéma d'annotation XML de type MUC [Hirschman & Chinchor, 1997] simplifié. Ne souhaitant travailler que sur les occurrences du pronom *it* impersonnelles ou anaphoriques, un balisage simple suffit :

> *<IMP ID="4">It</IMP>was shown that <EXP ID="5">a cis-acting catabolite-responsive element CRE sequence</EXP> located 64 bp downstream of the transcription-start site mediated catabolite repression of the dra-nupC-pdp operon as <EXP ID="3"><REF SRC="5"/>it</EXP> does for many other B. subtilis genes.*

Les occurrences impersonnelles sont marquées par la balise *IMP* et les occurrences anaphoriques ainsi que les antécédents par la balise *EXP*, la balise *REF* introduit le type *SRC*, un entier qui pointe vers l'antécédent du pronom anaphorique, dans la phrase précédente, le pronom d'indice 3 est anaphorique et son antécédent est le SN d'indice 5.

Corpus Transcript : Le corpus sur lequel nous avons évalué notre système est le corpus Transcript. Ce corpus, initialement créé pour le projet Caderige[51] : [Alphonse *et al.*, 2004], est composé de 2209 résumés d'articles scientifiques de génomique extrait de la base `Medline` avec la requête : *bacillus subtilis, transcription*[52]. 697 occurrences du pronom *it* ont été identifiées dans Transcript. Deux annotateurs ont étiqueté séparément les relations anaphoriques et impersonnelles avec le schéma du paragraphe précédent. Le taux d'accord inter-annotateur est de 92.4%. Le tableau suivant résume les caractéristiques de ce corpus.

Les ressources linguistiques et logicielles

Ce travail est lié au projet européen *ALVIS*[53]. Pour déterminer les valeurs des attributs nous avons analysé nos corpus au moyen de la plateforme linguistique *OGMIOS* réalisée lors de ce projet [Derivière *et al.*, 2006], [Hamon & Nazarenko, 2008]. Cette plateforme produit une analyse du corpus sous format XML, conforme à une DTD prédéfinie pour notre système de résolution.

[51] `http://caderige.imag.fr/`
[52] Le corpus Transcript est une sous-partie du corpus Génétique.
[53] Références du projet : IST-1-002068-STP

```
<ABSTRACT>
<SENTENCE>
<header>This article is about Mutational Analysis of Conserved Residues in the Putative
DNA-Binding Domain of the Response Regulator Spo0A of Bacillus subtilis .</header>
</SENTENCE>
 <WORDS>
  <word>This</word><tag>DT</tag><lemm>this</lemm>
  <word>article</word><tag>NN</tag><lemm>article</lemm>
  <word>is</word><tag>VBZ</tag><lemm>be</lemm>
  <word>about</word><tag>RB</tag><lemm>about</lemm>
  <word>Mutational</word><tag>JJ</tag><lemm>mutational</lemm>
  <word>Analysis</word><tag>NN</tag><lemm>analysis</lemm>
  <word>of</word><tag>IN</tag><lemm>of</lemm>
                                  ° ° °
 </WORDS>
 <NAMED_ENTITIES>
  <named_entity>Spo0A</named_entity><named_entity_type>gene</named_entity_type>
  <named_entity>Bacillus subtilis</named_entity><named_entity_type>species</named_entity_type>
 </NAMED_ENTITIES>
 <TERMS>
  <term>Analysis</term>
  <term>Bacillus subtilis</term>
 </TERMS>
<SYNTACTIC_CONSTITUANTS>
<constituant>S <constituant>NP This article </constituant> <constituant>VP is
<constituant>PP about ....  </constituant> . </constituant>
</SYNTACTIC_CONSTITUANTS>
 <SYNTACTIC_RELATIONS>
  <relation_type>'COMP:V-N'(4)</relation_type>
  <word1>is</word1><word2>Analysis</word2>
  <position_word1>3</position_word1><position_word2>6</position_word2>
                                  ° ° °
 </SYNTACTIC_RELATIONS>
 <SENTENCE>The Spo0A protein of Bacillus subtilis is a DNA-binding protein that is required
for the expression of genes involved in the initiation of sporulation . </SENTENCE>
 <WORDS>
                                  ° ° °
```

FIG. 6.2 – Un texte annoté au format XML d'entrée pour notre système

Hormis le segmenteur et l'étiqueteur d'entités nommées, nous avons paramétré la plateforme de manière à utiliser différents outils selon nos corpus. La segmentation est effectuée au moyen d'un algorithme à base d'expressions régulières. Cette étape est cruciale pour notre système mais son calcul est très bruité sur le corpus de génomique (du fait de la présence de mesures, d'équations, de listes etc...). Nous avons pris le parti de corriger manuellement cette annotation pour les deux corpus plutôt que d'utiliser un outil spécifique à chacun. C'est l'unique correction apportée aux données d'entrée du système. L'étiqueteur d'entités nommées par défaut de la plateforme se nomme TagEN [Berroyer, 2004] et emploie les techniques classiques pour ce type d'annotation, une grammaire à base de transducteurs complétée par des ressources

linguistiques. Cet étiqueteur spécifie pour chaque entité nommée la classe sémantique à laquelle elle appartient. Lorsque nous avons procédé à nos expériences, seules les connaissances nécessaires au traitement des textes de génomique étaient disponibles.

Préparation du corpus MARS Le style d'écriture des textes du corpus MARS est celui des manuels d'utilisation de matériels ou de logiciels informatiques donc avec un phrasé proche de la langue courante. Nous avons retenu des outils génériques pour les annoter. L'annotation morpho-syntaxique a été calculée par le TreeTagger[54], l'analyse syntaxique en dépendances et en constituants par le Link Parser[55] et les termes ont été calculés sur le corpus avec YaTeA [Aubin & Hamon, 2006][56]. Nous avons gardé l'ensemble des verbes indicatifs de Mitkov.

Préparation du corpus Transcript Le vocabulaire des résumés d'articles de génomique est bien plus spécialisé que celui des manuels techniques et réclame un traitement adapté. L'étiquetage morpho-syntaxique a été réalisé avec le Genia Tagger[57], les analyses syntaxiques en constituants et en dépendances ont été réalisées par BioLG, une version du Link Parser adaptée à notre domaine [Pyysalo *et al.*, 2006]. Les termes n'ont pas été calculés automatiquement sur notre corpus mais projetés à partir de ressources préexistantes du domaine, Gene Ontology et le MeSH [Derivière *et al.*, 2006] pour avoir une couverture plus importante. Enfin, nous avons complété les classes de verbes indicatifs de Mitkov [Mitkov, 2002] à partir d'un corpus d'entraînement.

6.2 Première expérience : évaluation du système de référence *Bio-MARS*

Dans cette première expérience notre objectif est de valider notre implémentation du système MARS, que nous appelons Bio-MARS, un système qui nous servira de système de référence dans la section suivante. Pour cela nous avons comparé ses performances sur le corpus de références MARS avec les performances publiées de l'implémentation de [Mitkov *et al.*, 2002].

Nous avons choisi d'implémenter notre propre version du système MARS car la première version du système est indisponible (voir la section 2.2.3 pour une description de ce système). Seule la deuxième version du système MARS est disponible (voir [Mitkov *et al.*, 2002] pour une description de ce système). Mais cette deuxième version du système ne convient pas comme système étalon. Elle intègre des connaissances linguistiques complexes comme le parallélisme syntaxique ou l'identification des entités animées dans la résolution. Notre objectif étant de mesurer l'apport de ces informations dans la résolution, il nous faut comparer le comportement d'un système n'exploitant que des indices de surface avec un système plus complexe intégrant ces connaissances linguistiques.

Nous aurions pu modifier le code de la deuxième version du système MARS pour n'utiliser que les indices de surface du système mais le coût, en termes de temps et d'effort, nous a paru

[54] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[55] http://www.link.cs.cmu.edu/link/
[56] http://www-lipn.univ-paris13.fr/~hamon/YaTeA/index.html
[57] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

plus important que celui d'une réimplémentation de la première version du système. Le système MARS d'origine est très bien documenté dans [Mitkov, 2002] et nous avons pu réutiliser le code de ce système pour implémenter notre classifieur bayésien.

6.2.1 Protocole expérimental

Les systèmes que nous avons testés dans cette section ne sont pas des systèmes d'apprentissage, il est donc nul besoin de procéder à une validation croisée pour juger la qualité de ces derniers. Nous rapportons leurs performances sur chaque document du corpus, puis sur la totalité du corpus. Les paramètres du filtre des pronoms impersonnels et les patrons de collocations ont été calculés sur l'ensemble du corpus.

6.2.2 Une implémentation du système MARS satisfaisante

Lors de cette expérience nous mettons en compétition six systèmes. Les deux premiers systèmes sont les systèmes que nous souhaitons comparer : le système MARS dans sa version d'origine et notre implémentation Bio-MARS. Les quatre autres systèmes servent de systèmes étalons. Le système *Aléatoire* choisit un antécédent au hasard dans la liste des candidats. Le système *Premier GN* sélectionne toujours comme antécédent le premier GN de la phrase contenant le pronom ou de la phrase précédant si le pronom est le premier GN de la phrase. Les systèmes *MAX* et *MAX-Origine* ont accès aux relations anaphoriques annotées du corpus et ils reconnaissent toujours l'antécédent dans la liste de candidats lorsqu'il s'y trouve. Le système *MAX* sélectionne l'antécédent dans une liste de candidats identique à celle du système Bio-MARS, le système *MAX-Origine* dans une liste de candidats identique à celle du système MARS. En raison des erreurs contenues dans les analyses syntaxiques en constituants des systèmes Bio-MARS et MARS à partir desquelles la liste des candidats est calculée, certains GN candidats ne sont identifiés que partiellement ou font défaut[58]. Les performances des systèmes *MAX* et *MAX-Origine* ne peuvent donc pas atteindre 100% et les deux dernières colonnes donnent les scores maximums possibles pour leur résolution.

Une comparaison *stricto sensu* des performances des deux systèmes a peu de sens. Le système MARS n'est pas restreint à la résolution du pronom *it* et les performances pour ce pronom seul ne sont pas isolées dans les résultats publiés. De plus les deux systèmes reposent sur des analyseurs syntaxiques différents et, Bio-MARS n'ayant pas été spécialisé pour le corpus MARS, il ne dispose pas des informations *Référence immédiate* et *Instruction séquentielle*. La confrontation des résultats du système MARS d'origine avec les performances de Bio-MARS obtenues sur le corpus sert seulement à mettre en évidence le bon comportement de notre réimplémentation du système MARS.

L'écart inférieur à 15% entre le taux de succès partiel du système Bio-MARS et du système MARS sur la totalité du corpus (voir le tableau 6.3) est suffisant pour admettre que le comportement de notre implémentation est comparable à celui de la version d'origine. Les différences observées sur les documents proviennent essentiellement de la variation des performances des

[58] Le link parser et BioLG n'analysent pas les phrases de plus de 70 mots ou ne contenant pas de verbe. Lorsque l'analyse est manquante nous créons la liste des candidats grâce à des heuristiques reposant uniquement sur le POS-Tagging.

analyseurs syntaxiques. Par exemple, notre analyseur syntaxique analyse très partiellement le document *Access* avec une performance du système *MAX* limitée à 52.88 pour le taux de succès strict et 70.19% pour le partiel, alors que l'analyseur syntaxique du système MARS produit une meilleure analyse de ce document avec des performances du système *MAX* de la version d'origine égales à 73.88% et 96.18% en taux de succès strict et partiel respectivement (ces mesures ont été définies dans la section 2.1.5). En conséquence, les performances du système Bio-MARS sont inférieures à celles du système MARS sur ce document. Le rapport s'inverse lorsque Bio-MARS bénéficie d'une meilleure analyse que MARS comme c'est le cas pour le document *Winhelp*. Notre analyseur retrouve partiellement 100% des candidats contre 87.50% pour l'analyseur de MARS.

	Aléatoire		Premier GN		Bio-MARS		MARS		*MAX*		*MAX-Origine*	
	Strict	*P.*	*Strict*	*Partiel*	*Strict*	*Partiel*	*S.*	*Partiel*	*Strict*	*Partiel*	*Strict*	*Partiel*
Access	5.76	-	5.77	7.69	11.54	11.54	-	33.33	57.69	70.19	73.88	96.18
Aiwa	11.11	-	22.22	22.22	55.56	55.56	-	35.48	77.78	77.78	81.43	95.71
Cdrom	11.32	-	22.64	30.19	39.62	50.94	-	53.84	73.58	84.91	78.31	95.18
Hinari	0	-	14.29	14.29	42.86	42.86	-	-	71.43	85.71	-	-
Panason	8.33	-	33.33	33.33	50.00	58.33	-	-	91.67	91.67	-	-
Urban	11.11	-	55.55	55.55	44.44	66.67	-	-	77.78	100	-	-
Winhelp	6.45	-	16.13	19.35	32.26	38.71	-	33.32	77.42	100	81.25	87.50
Corpus	7.69	-	16	19	28.21	32.91	-	46	69	81	-	-

TAB. 6.3 – Détail des taux de succès des systèmes MARS (version d'origine) et Bio-MARS sur le corpus MARS

6.3 Deuxième expérience : évaluation du classifieur bayésien

Dans cette section nous estimons puis analysons les performances de notre classifieur bayésien pour la résolution des anaphores sur les corpus Transcript et MARS en vue de répondre à notre question initiale : un système de résolution des anaphores qui intègre l'incertitude des annotations en paramètres de son mécanisme d'inférence améliore-t-il ses scores ?

6.3.1 Le protocole d'expérimentation

Nos deux corpus sont de petite taille donc nous avons, pour chacun, estimé la moyenne des performances des systèmes en concurrence avec une validation croisée où le nombre d'itérations et le pourcentage de documents d'entraînement diffèrent selon les corpus. Le corpus MARS est constitué d'un petit nombre de documents qui contiennent un nombre très variable de pronoms. En raison de ces caractéristiques nous avons opté pour une validation croisée de 7 itérations où les 6 premiers documents sont réservés à l'apprentissage et le dernier au test. Le corpus Transcript a des caractéristiques différentes : il est composé d'un nombre important de documents qui contiennent chacun peu de pronoms. Pour ce corpus nous préférons une validation croisée

avec 10 itérations où les deux tiers du corpus sont affectés au corpus d'entraînement. Le tiers restant est réservé pour le corpus de test. En raison du nombre important d'erreurs obtenues sur la totalité des itérations, nous avons limité mais détaillé l'analyse des erreurs de nos systèmes aux erreurs de la première itération.

Nous avons résolu les anaphores avec sept systèmes différents sur le corpus Transcript. Les quatre systèmes *Aléatoire*, *Premier GN*, *MAX* et *Bio-MARS* servent de systèmes de comparaison. Trois autres systèmes ont été retenus pour tester différentes configurations du modèle bayésien. Le système *Naïf-MARS* exploite les mêmes attributs que le système *Bio-MARS* mais la décision finale est prise par un classifieur bayésien naïf et non par le module de score original. Le sixième système est notre classifieur bayésien *CB* reposant sur un réseau bayésien et exposé dans la section précédente. Le dernier système est le classifieur bayésien naïf *Naïf-CB* associé au *CB i.e.* un classifieur qui possède les mêmes variables aléatoires que le système *CB* mais où elles sont toutes supposées indépendantes.

Sur le corpus MARS nous avons réduit le nombre de systèmes à six. Nous prenons de nouveau les systèmes *Aléatoire*, *Premier GN*, *MAX*, *Bio-MARS* comme systèmes de comparaison et nous évaluons les résultats des classifieurs *Naïf-MARS* et *CB* sur un corpus d'un domaine différent de la génomique[59].

Pour ces expériences nous avons modifié la méthode de calcul de la métrique du taux de succès partiel employée par Mitkov. Pour cet auteur un antécédent est partiellement reconnu lorsque seule une partie incomplète contenant la tête de l'antécédent à été annotée. Nous ajoutons une contrainte supplémentaire : la partie annotée doit toujours contenir la tête de l'antécédent et pouvoir être substituée au pronom anaphorique sans incohérence. Considérons la phrase :

> *[beta-Galactosidase expression from the spl-lacZ fusion]$_1$ was silent during vegetative growth and was not DNA damage inducible, but [it]$_1$ was activated at morphological stage III...*

notre système annote uniquement *beta-Galactosidase expression*, un candidat qui peut être substitué au pronom. Dans cette seconde phrase :

> *The larger mother cell engulf [the smaller forespore]$_1$, then nurtures [it]$_1$ and eventually, lyses to release a dormant, environementally resistant spore.*

le mauvais candidat *The larger mother cell engulf the smaller forespore* ne pouvant se substituer au pronom sans glissement de sens il est rejeté, même s'il contient l'antécédent entièrement.

6.3.2 Corpus MARS : résultats et discussions

En appliquant les trois systèmes sur le corpus MARS notre objectif est double. La confrontation des scores des systèmes *Bio-MARS* et *Naïf-MARS* permet d'apprécier l'apport du changement de représentation et de l'apprentissage dans la décision du système. Les attributs employés par les deux systèmes étant rigoureusement identiques dans leurs valeurs et leur calcul, seul varie le mécanisme d'inférence. L'opposition des taux de succès des systèmes *Naïf-MARS* et *CB* nous apprend, quant à elle, l'apport des attributs syntaxiques. Le mécanisme d'inférence des

[59] Nous calculons les performances du système *Naïf-CB* afin de les comparer avec celles du système *CB* dans le but d'établir l'éfficacité du mécanisme de correction du réseau bayésien. Nous avons obtenu ces résultats sur le corpus Transcript (voir la section 6.3.3) et jugé inutile de les confirmer sur le corpus MARS.

deux systèmes repose sur une base identique, un classifieur bayésien, seules varient les informations à disposition des systèmes. Le système *Naïf-MARS* dispose uniquement des indices de surface du système MARS alors que le système *CB* bénéficie en principe de tous les indices décrits dans la section 6.1.2. A noter cependant que l'annotateur d'entités nommées de la plateforme *Ogmios* n'a pas été spécialisé pour le domaine technique. Les entités nommées retrouvées ont été peu déterminantes et très peu nombreuses. Dans ce corpus seules les dates, les noms de personne, les lieux et les adresses internet ont été annotés alors que les noms de logiciels ou de matériels éléctroniques, bien plus significatifs, ont été ignorés. En conséquence, le système *CB* est dans l'impossibilité de calculer les indices reposant sur les entités nommées et leur classes sémantiques. Il dispose uniquement des informations syntaxiques en plus des informations de *Naïf-MARS*. Le tableau 6.8 récapitule les taux de succès de chaque système par document.

Un filtre moins performant pour les pronoms impersonnels

Un score minimal : En raison du changement de domaine, nous nous attendions à une moindre performance du filtre des pronoms impersonnels. Et, en effet, on observe une exactitude de 87.5% sur le corpus entier (voir le tableau 6.4). Toutefois, il s'agit d'un score minimal car des imperfections de l'annotation des données de test et des erreurs dans l'implémentation de notre système dégradent les performances. Faute de temps, nous n'avons pu corriger ces erreurs et renouveler l'expérience. Nous considérons qu'après la correction de ces erreurs le système obtiendrait un score de 92.7%[60]. (voir le tableau 6.5 pour la typologie des erreurs et le tableau 5 de l'annexe 2 pour le détail des phrases).

	Vrais Négatifs	Vrais Positifs	Faux Négatifs	Faux Positifs	Exactitude
Access	96	10	10	8	85.5%
Aiwa	16	0	0	2	88.9%
Cdrom	52	6	4	1	92.1%
Hinari	7	3	1	0	90.9%
Panason	11	9	2	1	87.0%
Urban	9	2	4	0	73.3%
Winhelp	29	2	1	2	91.1%
Corpus	220	32	22	14	87.5%

TAB. 6.4 – Taux de succès du filtre de pronoms impersonnels sur le corpus MARS

Les erreurs du système pour les 15 pronoms du tableau 6.5 sont à imputer à des erreurs humaines ou à des erreurs d'implémentation et peuvent être corrigées facilement. La première erreur est due à une annotation que nous contestons :

> *If your software is only usable through a graphical interface then [it] can be very hard to make it usable for someone who can't see.*

A nos yeux le pronom est impersonnel et correctement étiqueté par notre système.

[60] Le score de 92.7% a été obtenu grâce à un calcul d'exactitude identique à celui du premier score, *i.e.* 87.5%, où nous avons supposé l'ensemble des erreurs corrigées.

Types des erreurs	Nombre d'erreurs
Annotation contestée	1
Anaphore causale annotée impersonnelle	3
Anaphore verbale annotée impersonnelle	4
Mauvais traitement d'un caractère non-alphanumérique	4
Erreurs d'implémentation diverses	3

TAB. 6.5 – Détail des erreurs humaines corrigeables

Les 7 erreurs suivantes sont dues à une mauvaise préparation du corpus. Le corpus annoté dont nous avons hérité ne distinguait pas les pronoms anaphoriques clausaux[61] ou verbaux[62] des pronoms impersonnels. Lors de notre préparation du corpus annoté destiné à être analysé par notre système, nous avons par erreur étiqueté 7 pronoms anaphoriques comme impersonnels.

Les 7 dernières erreurs sont dues à des erreurs dans l'implémentation de notre système. 4 d'entre elles proviennent d'un mauvais traitement d'un caractère non-alphanumérique. Prenons la phrase :

> *If you can read from the drive but cannot mount it, first verify that you compiled in ISO-9660 file system support by reading/proc/filesystems, as described previously.*

cette séquence est reconnue par une règle de Paice. Nos règles acceptent les virgules pour permettre la reconnaissance des appositions. La virgule qui suit immédiatement le pronom est, dans notre cas, confondue avec la première virgule d'une apposition. Nos règles doivent être corrigées pour refuser ce type de séquences. Les 3 dernières erreurs sont dues à des anomalies dans le déroulement du programme.

Causes	Nombre
Mauvais paramétrage	7
Erreur d'un automate	1
Descriptions non discriminantes	2

TAB. 6.6 – Détail des Faux Positifs

Analyse des Faux Positifs : A l'exception des faux positifs dus à un mauvais paramétrage du système, la typologie des faux positifs du système est similaire à celle des faux positifs du corpus Génétique, nous détaillons les erreurs dans le tableau 6.6 et donnons les phrases complètes dans le tableau 3 de l'annexe 2.

7 FP se produisent sur un mauvais paramétrage du système. Le 1^{er} document contient la moitié des pronoms. Lorsque ce texte n'est pas présent dans le corpus d'apprentissage, le corpus contient peu d'exemples d'apprentissage et les paramètres appris sont moins fiables. Le

[61] Ex. *In principle [it should be possible to put together a complete, usable Linux system for a visually impaired person for about $500]$_1$ (cheap and nasty PC + sound card).[...] I doubt [it]$_1$ would work in practice because the software speech synthesisers available for Linux aren't yet sufficiently good.*

[62] Ex. *If someone [gets this to work properly]$_1$, please send me the details of how you did [it]$_1$.*

système étiquette 3 pronoms comme impersonnels alors qu'ils sont anaphoriques. Il les étiquette correctement si l'on apprend les paramètres sur l'ensemble du corpus. Les séquences de ces pronoms finissent par les délimiteurs peu fréquents *how, which, when* et contiennent des adjectifs/verbes/noms connus. Les 4 autres pronoms anaphoriques classés impersonnels par le système apparaissent dans des séquences terminées par le délimiteur *to* et d'une longueur supérieure à 5 mots entre le pronom et le délimiteur. Durant l'étape d'apprentissage l'imprécision de nos transducteurs a conduit le système à sur-évaluer le poids des séquences longues. Notre corpus d'apprentissage est composé de quelques séquences où le pronom est impersonnel et qui contiennent deux occurences de la préposition *to*. Les transducteurs, qui sont imprécis, ignorent la première occurence de la préposition et choisissent la seconde comme délimiteur de la séquence. La valeur de l'attribut `Longueur_Pronom_Delimiteur` des vecteurs qui décrivent ces exemples d'apprentissage est donc incorrecte, car plus grande que celle attendue, ce qui fausse l'apprentissage de la probabilité *a priori*. Le système, s'appuyant sur ces exemples d'apprentissages bruités, suppose qu'un pronom impersonnel a une forte probabilité d'apparaître dans une séquence longue et il classe les 4 pronoms anaphoriques du corpus de test comme impersonnels.

L'erreur suivante est due à une erreur d'un automate qui reconnaît la séquence :

> *However, there seem to be interactions between xpuff and the window manager which could make it difficult to use.*

Cette séquence est reconnue à la fois par notre automate et une règle de Paice, le pronom est donc malheureusement annoté impersonnel.

Les 2 derniers FP sont des séquences terminées par le délimiteur *to* dont les descriptions sont non discriminantes sur notre corpus d'apprentissage.

Causes	Nombre
Absence de ressource	5
Forme interrogative non-gérée	1
Délimiteur implicite/absent	2
Délimiteur non-géré	2
Transducteur imprécis	1

TAB. 6.7 – Détail des Faux Négatifs

Analyse des Faux Négatifs : Les raisons qui permettent de comprendre les FN du système sont multiples. Nous les résumons dans le tableau 6.7 et donnons à nouveau les phrases dans le tableau 4 de l'annexe 2.

En raison du changement de domaine 5 pronoms ont été mal reconnus car aucun des verbes/adjectifs/noms des séquences n'appartenaient à nos ressources[63]. Un pronom apparaît dans la question

What does it mean when I get a kernel message from the IDE CD-ROM driver like "hdxx :[...]

[63] Ex. *There are also others which it is worth researching which cover computer use more generally.*

et se voit mal étiqueté. Les formes interrogatives ne sont pas gérées par notre système car elles étaient absentes de notre corpus de génétique. 2 erreurs sont de nouveau dues à un délimiteur implicite ou absent de la séquence et 2 autres erreurs à des délimiteurs que nous n'avons pas modélisés dans nos automates car très peu présents dans le corpus. Il s'agit des délimiteurs *for* et *what*. La dernière erreur est due à un transducteur imprécis.

Bio-MARS, un lauréat inattendu

Pour cette expérience nous avons mis en concurrence 6 systèmes : *Aléatoire, Premier GN, Max, Bio-MARS, Naïf-MARS* et *CB*. Les scores de la résolution des anaphores des systèmes en compétition sont proches et en faveur du système *Bio-MARS* (voir le tableau 6.8). Les chiffres de notre expérience confirment le résultat inattendu de [Mitkov, 2002]. Les poids du système *Bio-MARS* déterminés par l'observation sont de très bonne qualité et rivalisent avec ceux appris automatiquement.

Les points forts et les points faibles de la stratégie du système *MARS* sur ce corpus ont déjà été mis en évidence par [Mitkov, 2002]. Les méthodes qui exploitent des connaissances linguistiques complexes échouent souvent sur les relations où l'antécédent et l'anaphore n'ont pas le même rôle syntaxique ou sémantique car ils privilégient en général le parallélisme syntaxique ou sémantique. L'approche pauvre en connaissances, qui ne dispose pas de ces connaissances, résoud mieux ce type d'anaphores en traquant par de multiples indices l'élément saillant qui est supposé être le meilleur candidat à l'antécédence. Cette stratégie a toutefois ses limites. Il est difficile de retrouver l'élement saillant dans une phrase à la construction syntaxique complexe sans connaissances syntaxiques et sémantiques. Nous serons aussi confrontés à cette difficulté lors de notre expérience sur le corpus Transcript (voir la section 6.3.3).

	Bio-MARS		NB_MARS		CB		*MAX*	
	Strict	*Partiel*	*Strict*	*Partiel*	*Strict*	*Partiel*	*Strict*	*Partiel*
Access	11.54	11.54	16.35	16.35	17.31	17.31	57.69	70.19
Aiwa	55.56	55.56	16.67	16.67	16.67	16.67	77.78	77.78
Cdrom	39.62	50.94	33.96	37.74	37.74	50.94	73.58	84.91
Hinari	42.86	42.86	14.29	14.29	14.29	14.29	71.43	85.71
Panason	50.00	58.33	58.33	66.67	50.00	58.33	91.67	91.67
Urban	44.44	66.67	55.56	66.67	55.56	55.56	88.89	100
Winhelp	32.26	38.71	22.58	29.03	19.35	35.48	77.42	100
Corpus	28.21	32.91	24.79	27.35	25.21	31.20	68.80	80.77

TAB. 6.8 – Détail des taux de succès des systèmes sur le corpus MARS

Bio-MARS vs. Naïf-MARS

Un apprentissage non profitable des paramètres Au vu des scores du tableau 6.8, le système *Bio-MARS* réalise une meilleure classification que le système *Naïf-MARS*. Un résultat surprenant, car même si les deux systèmes emploient les mêmes attributs, la représentation de l'incertitude du système *Bio-MARS* est plus simple que celle du système *Naïf-MARS*. Les poids

des attributs du système *Bio-MARS* ne sont justifiés que par l'observation alors que le système *Naïf-MARS* bénéfécie d'une étape d'apprentissage pour trouver les meilleurs poids pour le corpus. L'analyse du tableau 6.9 montre que la majorité des erreurs qui sont propres au système *Naïf-MARS* proviennent d'un mauvais apprentissage du poids des attributs. Pour simplifier l'exposé du détail des erreurs nous avons limité notre étude aux anaphores strictement résolues par un des systèmes et incorrectement résolues par l'autre système.

Détail des erreurs des systèmes *Bio-MARS* et *Naïf-MARS*		
151 erreurs communes aux 2 systèmes		
Causes des décisions divergentes entre les deux systèmes	Antécédents manqués par *Bio-MARS* et strictement retrouvé par *Naïf-MARS*	Antécédents manqués par *Naïf-MARS* et strictement retrouvé par *Bio-MARS*
Mauvaise évaluation du poids de la répétition d'un candidat	12	15
Sous-évaluation du poids des patrons de collocations, des candidats en tête de section et de la répétition d'un candidat	1	3
Représentation probabiliste de l'incertitude défavorable	4	7
Total	17	25
	Bio-MARS : 168 erreurs	Naïf-MARS : 176 erreurs

TAB. 6.9 – Détail des erreurs communes et propres aux systèmes *Bio-MARS* et *Naïf-MARS* sur le corpus MARS

Analyse détaillée des erreurs Les différences de performance observées entre les deux systèmes viennent essentiellement d'une mauvaise évaluation du poids attribué aux candidats qui sont répétés plus d'une fois dans le document (*i.e.* la valeur *plus* de l'attribut *SN_Repete*. Une caractéristique des documents et le bruit de l'analyse syntaxique ont faussé l'apprentissage de la probabilité de cet attribut. Les documents de notre corpus sont volumineux et répètent souvent certains GN qui ne sont pas nécessairement repris anaphoriquement. L'imperfection de l'analyse syntaxique ajoute encore à l'imprécision de cet attribut. Certains syntagmes ne sont pas analysés entièrement et se limitent au seul déterminant *the* ou *this*. Lorsque ces syntagmes sont en position de candidat, le système va calculer le nombre de fois où ils sont répétés dans le document pour déterminer les fréquences nécessaires aux calculs des probabilités conditionnelles *a priori* de l'attribut `SN_Repete`. La répétition d'un candidat est calculée par une simple comparaison de chaînes de caractères. Si le candidat se réduit au seul déterminant *the*, il est retrouvé un grand nombre de fois dans le document et le plus souvent il n'est pas antécédent. Le système calcule alors de mauvaises valeurs pour les deux probabilités suivantes : *P(SN_Repete=Plus|Antecedent)=21.8* contre *P(SN_Repete=Plus|NonAntecedent)=27.3* : alors que le système *Bio-MARS* avantage fortement un candidat qui a été répété deux fois ou plus dans le corpus, le système *Naïf-MARS* les pénalisent. Ce comportement contraire explique à lui

seul 15 erreurs commises par le système *Naïf-MARS* alors que le système *Bio-MARS* retrouve l'antécédent. Dans la phrase :

[Kenneth Albanowski "kjahds@kjahds.com"]$_1$ provided [the patch]$_2$ needed for the Brailloterm and information about [it]$_2$.

l'antécédent obtient un score de 4 car il est dans la même phrase que le pronom, il est indéfini et il est répété 3 fois dans le document. Le système *Bio-MARS* le place en première position. Le système *Naïf-MARS* pour qui la répétition est une information négative déclasse l'antécédent à la deuxième position avec une probabilité de 27.0% et lui préfère le premier candidat avec une probabilité de 49.3% [64]. Bien sûr, il arrive que le mauvais poids de cet attribut avantage le système *Naïf-MARS*, comme dans la séquence

[The device file]$_1$ can be created using : # mknod /dev/mcdx0b200. If you recently bought [a Mitsumi CD-ROM drive]$_2$, don't assume that it should use this kernel driver.

où le score attribué pour la répétition du premier candidat permet d'atténuer sa mauvaise position et trompe le système *Bio-MARS* alors que le système *Naïf-MARS* lui préfère le candidat non filtré de la phrase du pronom avec une probabilité de 15.3%. Mais cette situation se présente peu dans ce corpus avec seulement 12 cas d'erreurs où le mauvais poids de cet attribut avantage le système *Naïf-MARS*[65].

Trois autres erreurs commises par le système *Naïf-MARS* et correctement résolues par le système *Bio-MARS* proviennent d'une sous-évaluation du poids des patrons de collocations, des candidats en tête de section et de la répétition d'un candidat[66]. L'antécédent de la phrase :

If [Quick View]$_1$ is not available, either [your program]$_2$ doesn't support [it]$_1$, or [it]$_1$ is not installed on your computer.

dont une occurrence apparaît dans un titre de section est indéfini et en première position mais aussi précédé par *if* donc en position prépositionnelle pour le système *Bio-MARS*[67] qui lui donne en conséquence un score de 3. Le candidat *your program* est un syntagme défini dans la même phrase et se voit attribuer un score de 2. Le système *Naïf-MARS* qui sous-évalue le poids de la répétition et pénalise un candidat apparaissant dans la tête de section (avec des probabilités *P(SN_Tete_Section = TeteSection|Antecedent) = 0.027* et *P(SN_Tete_Section = TeteSection|Antecedent) = 0.038)* inverse l'ordre des candidats et choisit *your program* avec une probabilité de 30.7% contre une probabilité de 30.0% pour l'antécédent. La sous-évaluation de ces attributs n'avantage le système *Naïf-MARS* que pour une unique anaphore.

Les 7 dernières erreurs produites par le système *Naïf-MARS* et correctement résolues par

[64] Le genre du corpus, ici des documents techniques, a aussi une incidence dans le nombre d'erreurs du système *Naïf-MARS* causées par cet attribut. Le deuxième document dans lequel l'écart entre les deux systèmes est frappant, est un manuel technique qui décrit l'installation d'un téléviseur. Certaines instructions sont répétées à plusieurs reprises dans le document et l'une d'elles contient un pronom anaphorique : `2. Turn on the TV and set it for VCR reception.` Cette instruction apparaît six fois dans le document et pour chacune d'elles, l'antécédent a déjà été répété à de multiples reprises dans le document. Seul le système *Bio-MARS* retrouve l'antécédent pour toutes les répétitions.

[65] Dans le premier document, qui a été assez mal analysé par le Link Parser, le système *Bio-MARS* commet 5 erreurs sur des candidats tronqués comme *the*, *this*, *all*, en raison de leur répétition.

[66] Il s'agit ici de la valeur *One* de l'attribut qui discrimine correctement les antécédents mais dont le poids est insuffisant avec *P(SN_Repete=Un|Antecedent)=10.3* contre *P(SN_Repete=Un|NonAntecedent)=9.6.*

[67] Notre système détermine si un syntagme est en position prépositionnelle en vérifiant l'étiquette du mot qui le précède. Le *TreeTagger*, qui a analysé cette séquence, utilise une unique étiquette *IN* pour les prépositions et les conjonctions de subordination.

le système *Bio-MARS* proviennent d'une représentation de l'incertitude défavorable au système *Naïf-MARS* sur ce corpus. La représentation probabiliste est une représentation plus fine de l'incertitude et les cas d'égalité entre plusieurs candidats sont plus rares pour le système *Naïf-MARS*. Malheureusement sa préférence est érronée alors que le système *Bio-MARS*, qui "doute" facilement entre deux candidats, choisit avec raison le candidat le plus récent comme le montre l'exemple suivant. Les candidats 1 et 2 de la phrase :

> *[Serious injury]$_1$ to an individual, and damage to [the television]$_2$ may result if [it]$_2$ should fall.*

sont en concurrence pour la première position de la liste avec un score 2, *Serious injury* parce qu'il est le premier syntagme de la phrase et indéfini, *the television* parce qu'il est défini. Comme aucun autre critère d'arbitrage ne les distingue, le système *Bio-MARS* choisit le candidat de la liste le plus récent *i.e. the television*. Pour le système *Naïf-MARS*, l'attribut de la première position a un poids suffisamment important pour distinguer le candidat *Serious injury* de ses concurrents avec une probabilité de 49.3% (contre 30.7% pour l'antécédent). Le système *Naïf-MARS* choisit correctement l'antécédent pour 4 anaphores seulement alors que ce critère d'arbitrage indique le mauvais candidat au système *Bio-MARS*.

Devons-nous renoncer à notre représentation de l'incertitude ? Les analyses qui précèdent mettent en doute l'intérêt d'une représentation précise de l'incertitude et de l'apprentissage, qui constituait l'une de nos hypothèses de travail. Mais avant de la condamner, nous devons garder en mémoire que la taille du corpus d'apprentissage est relativement petite et que les paramètres ont été appris sur des données bruitées. De plus, les scores et les critères d'arbitrage ont été conçus sur et pour ce corpus. Nous verrons que le comportement du système *Bio-MARS* diffère sur un corpus d'un autre genre et d'un autre domaine dans l'expérience de la section suivante 6.3.3.

Naïf-MARS vs. CB

Un apport, faible mais visible, des informations syntaxiques Si les conclusions de la précédente section ont déçu nos espoirs dans le changement de représentation, l'avantage du système *CB* sur le système *Naïf-MARS* confirme l'intérêt de l'ajout des informations syntaxiques même bruitées dans un système de résolution. Même si l'écart entre les deux systèmes est très faible l'analyse détaillée des erreurs montre clairement que les informations syntaxiques sont utilisées par le système *CB* et changent la décision du classifieur dans certains cas. Ces cas sont peu nombreux car nos données ne sont que partiellement analysées avec un rôle grammatical des antécédents connus pour 30.6% d'entre eux, 24.8% pour la totalité des candidats du corpus et avec une précision relative de 64%[68].

Analyse détaillée des erreurs Le système *CB* réussit à résoudre correctement 3 anaphores grâce aux informations syntaxiques à sa disposition, alors que le système *Naïf-MARS*, qui en est dépourvu, sélectionne un mauvais candidat. Les deux systèmes divergent sur la résolution de l'anaphore de la phrase :

[68] Pour vérifier la précision nous avons vérifié les 100 premières relations syntaxiques *sujet* et *complément d'objet* trouvées sur le premier article et établi le pourcentage précédent.

Détail des erreurs des systèmes *Naïf-MARS CB*		
173 erreurs communes aux 2 systèmes		
Causes des décisions divergentes entre les deux systèmes	Antécédents manqués par *CB* et strictement retrouvés par *Naïf-MARS*	Antécédents manqués par *Naïf-MARS* et strictement retrouvés par *CB*
Apport de l'information syntaxique	0	3
Entité nommée erronée	1	0
Patron de collocation ambigü	1	0
	CB : 175 erreurs	Naïf-MARS : 176 erreurs

TAB. 6.10 – Détail des erreurs communes et propres aux systèmes *Naïf-MARS* et *CB* sur le corpus MARS

[It]$_1$ provides such features as a box around [the pointer]$_2$ which makes [it]$_2$ easier to locate.

Sans information syntaxique, le système *Naïf-MARS* choisit le premier candidat de la phrase alors que le système *CB* lui préfère le syntagme *the pointer* en raison du parallélisme syntaxique entre le candidat et le pronom qui sont tous deux compléments d'un verbe. Dans la phrase :

[Note]$_1$ that you should not eject [a disc]$_2$ while [it]$_2$ is mounted (this may or may not be possible depending on the type of drive).

en raison d'une erreur de prétraitement, le verbe *note* apparaît dans la liste des candidats[69] et, en vertu de sa position de premier syntagme de la phrase, il est choisi par le système *Naïf-MARS*. L'antécédent, bien que répété une fois dans le document et apparaissant dans un patron de collocation *NP-Verbe* obtient du système *Naïf-MARS* une probabilité de 20.2%, inférieure à la probabilité de l'autre candidat de 44.8%. Le système *CB* qui connaît en plus le rôle grammatical de l'antécédent et ignore celui du candidat, choisit correctement l'antécédent. Mentionnons toutefois que sans la connaissance du rôle grammatical de l'antécédent, le choix du système aurait été identique avec une perte de 2% uniquement sur la probabilité de l'antécédent. Le choix attendu est trouvé par un calcul plus "juste" des poids des différents attributs du *CB* grâce à sa structure. L'analyse en constituants de la dernière phrase de l'article :

[The sbpcd driver supports]$_1$ automatically ejecting [the CD]$_2$ when [it]$_2$ is unmounted.

est partiellement fausse. Le verbe *supports* est rattaché au syntagme *The sbpcd driver*, ce qui biaise le calcul du premier candidat. Le système *Naïf-MARS* choisit le premier syntagme de la phrase précédente. Pourtant le système *Naïf-MARS* crédite par défaut le candidat *The sbpcd driver* d'une probabilité de 30.7%, car ce candidat est défini et dans la même phrase que le pronom sans être répété plusieurs fois dans le document comme l'antécédent qui obtient une probabilité de 27.0%. Le système *CB* possède une description identique du candidat mais une description plus riche de l'antécédent. Ce dernier est complément du verbe *ejecting* et est retrouvé dans une collocation *Sujet-Verbe* identique à celle du pronom. Le système le choisit avec une probabilité de 99.2%. Notons qu'un calcul correct du premier syntagme n'aurait pas changé le résultat.

[69] Le *link Parser* analyse correctement la phrase mais, dans la version courante de notre système, le parenthésage rend impossible l'alignement entre l'analyse et la phrase du document. Le système en mode "panique" se sert de l'analyse morpho-syntaxique pour créér les candidats. Or cette analyse est erronée, le verbe *note* est étiqueté *nom* et donc inséré dans la liste des candidats.

Le système *Naïf-MARS* résout correctement deux anaphores alors que le système *CB* choisit un mauvais candidat. Mais les informations syntaxiques supplémentaires du *CB* ne sont pas en cause, une erreur d'annotation d'une entité nommée et un patron de collocation ambigü expliquent les erreurs du système. Le système *CB* choisit un mauvais candidat dans la phrase :

> *We recommend that a gap of at least 5 [cm]$_1$ is left all around [this TV]$_2$ even when [it]$_2$ is placed inside a cabinet or between shelves.*

en raison d'une conjonction de facteurs. Le premier candidat de la phrase est correctement filtré, ce qui oblige le système à choisir un candidat par défaut. Pour le système *CB* le candidat *cm* est un nom propre, ce qui l'avantage légèrement[70] par rapport à l'antécédent qui est répété 3 fois et dont le poids du pronom démonstratif est sous-évalué lors de l'apprentissage[71]. Le système *Naïf-MARS* préfère, avec une petite probabilité de 26.0%, le seul candidat défini de la phrase, *this TV*. L'autre erreur du *CB* se produit sur la première anaphore de la phrase :

> *You can open [the folder]$_1$ that contains [the file]$_2$ you want to move or copy, and then drag [it]$_2$ to the folder you want to put it in.*

un patron de collocations ambigü en est la cause. Un grand nombre d'attributs sont disponibles pour décrire le pronom, l'antécédent et le candidat. Ils sont tous deux définis, le candidat a été répété une fois dans le document, leurs rôles grammaticaux sont connus et ils ont tous deux un rôle syntaxique identique à celui du pronom[72]. Le candidat l'emporte sur l'antécédent car, comme le pronom, il a déjà été une fois complément d'objet du verbe *move*. Bien que cette collocation soit possible aussi pour l'antécédent (*move the file*), elle n'apparaît pas dans le corpus d'entraînement[73]. Soulignons, pour la défense du *CB*, que seule une singularité du corpus d'apprentissage protège le système *Naïf-MARS* d'une erreur identique. Le paramétrage appris sur le corpus composé des 6 premiers documents donne un poids négatif à la valeur `One` du nœud `SN_Repete`. Avec un poids positif normal pour cette valeur, le système *Naïf-MARS* choisit aussi le candidat et non l'antécédent, les descriptions étant identiques à l'exception de la valeur de cet attribut.

Conclusion provisoire sur l'apport des informations syntaxiques Au terme de cette expérience, nous ne disposons pas d'un nombre suffisamment important de résolutions divergentes des deux systèmes *Naïf-MARS* et *CB* pour conclure de manière définitive d'un apport positif des informations syntaxiques dans la résolution (même si on observe un écart plus marqué entre les deux systèmes lorsque l'on compare les résolutions partielles). Mais notre bilan provisoire y est favorable et montre au moins que l'ajout de ces informations bruitées ne dégrade pas les performances du système qui en bénéficie.

[70] Avec P(Nom_Propre=NomPropre|Antecedent,Inconnu)=0.224 contre P(Nom_Propre=NomPropre|NonAntecedent,Inconnu)=0.216.

[71] Le système *CB* corrige son erreur si les paramètres sont appris sur l'ensemble du corpus.

[72] Le candidat est engagé dans deux relations syntaxiques. Il est complément d'objet du verbe *open* et a donc le même rôle grammatical que le pronom, mais il est aussi sujet du verbe *contain*, ce sera la valeur retenue pour l'attribut `SN_Sujet`.

[73] Si une collocation identique pour l'antécédent vient à être connue, ce dernier n'est toujours pas choisi avec ce paramétrage, le poids du sujet est supérieur à celui du complément.

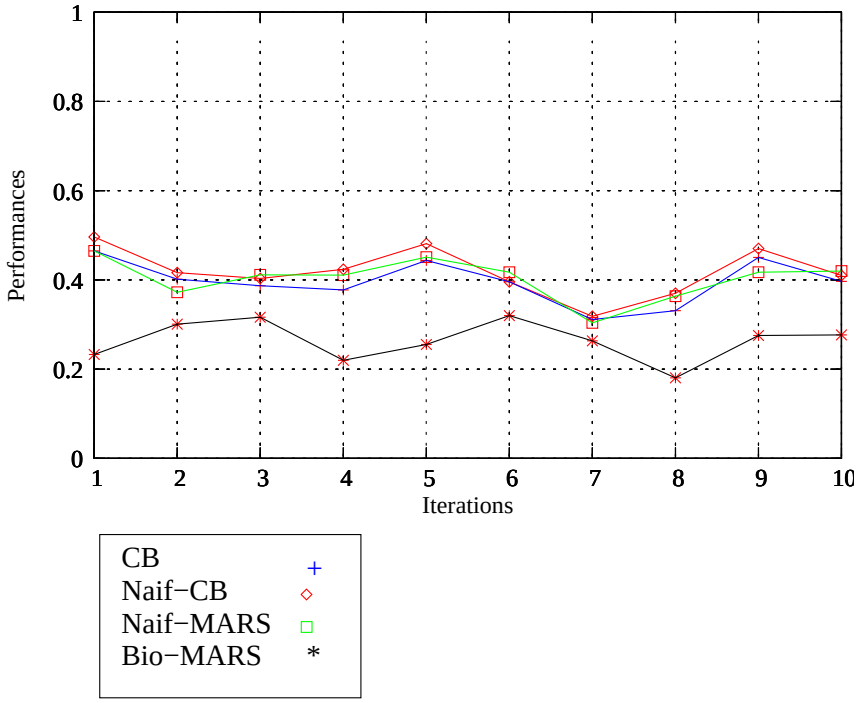

FIG. 6.3 – Détails des taux de succès stricts obtenus par les systèmes sur toutes les itérations

6.3.3 Corpus Transcript : résultats et discussions

Le tableau 6.11 résume la moyenne des taux de succès de nos sept systèmes et les figures 6.3 et 6.4 les scores obtenus par nos systèmes sur la totalité des itérations.

Une contre-performance du système *Bio-MARS*

Le premier fait marquant de ce tableau est une performance du système *Bio-MARS* inférieure sur notre corpus à celle obtenue par le système *Premier GN*. La différence étant fortement marquée entre les résultats du système *Bio-MARS* et ceux des autres systèmes, nous faisons l'économie d'une comparaison détaillée des erreurs de ce système avec celles des autres systèmes.

Dans la version probabiliste de *Bio-Mars*, *Naïf-MARS*, le paramétrage adapte les scores des attributs en accord avec les observations faites sur notre corpus et évite les erreurs dûes à la sur-évaluation des scores. De plus, le passage aux probabilités permet une représentation plus fine de l'incertitude, ce qui réduit le nombre de cas de candidats concurrents égaux et donc les erreurs d'arbritrage.

Analyse des erreurs du système *Bio-MARS* Nous indiquons dans le tableau 6.12 la typologie des erreurs qui sont propres au système *Bio-MARS* lors de la première itération.

Dans 50% des cas d'erreurs, l'élément saillant est mal reconnu en raison de la faillibilité des termes et du score trop important des patrons de collocations. La couverture de notre ressource terminologique est trop importante et distingue des mots qui, dans notre corpus, n'ont

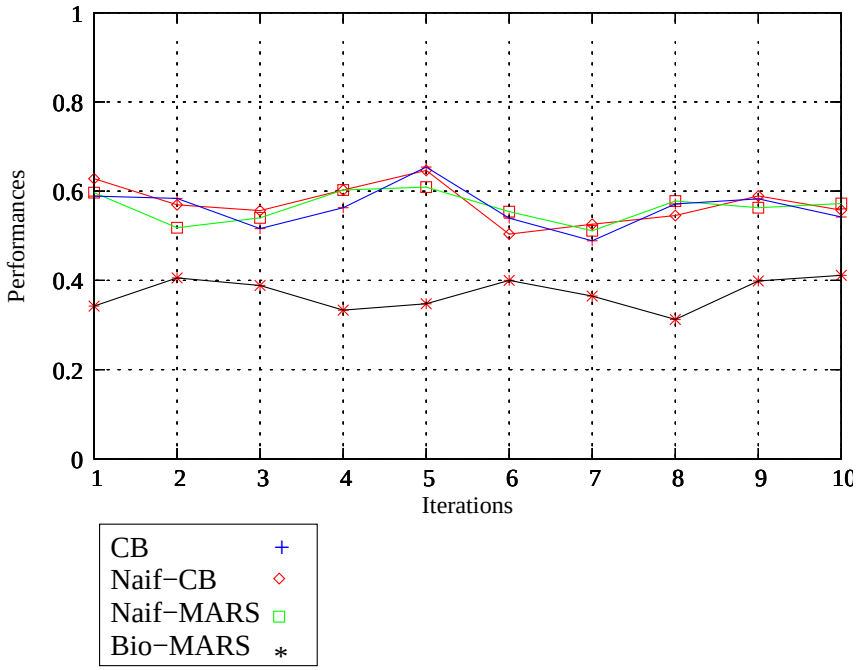

FIG. 6.4 – Détails des taux de succès partiels obtenus par les systèmes sur toutes les itérations

pas d'importance particulière comme *use* ou *work*. Les paramètres du *Naïf-MARS* montrent que P(Terme|NonAntecedent)=0.22 > P(Terme|Antecedent)=0.18. L'attribut des termes joue ici un rôle négatif dans la discrimination des antécédents. Le problème des patrons de collocations est différent. Ces derniers décrivent mieux les antécédents mais le score qui leur est attribué est trop élevé, si bien que la somme des indices marquant l'antécédent ne peut équilibrer celle d'un mauvais candidat retrouvé dans une collocation.

Pour 38% des erreurs, c'est la stratégie d'arbitrage entre deux candidats concurrents qui est à incriminer. Le choix de l'élément le plus récent propose des mauvais candidats et semble inadapté à notre corpus.

Les cas restants sont dûs aux erreurs du prétraitement : 3 antécédents sont filtrés car ils ont été mal annotés et 1 antécédent qui a été mal analysé syntaxiquement se limite au seul article *the*.

Des classifieurs aux performances similaires

Le second fait marquant du tableau, c'est la proximité contre-intuitive des performances des trois systèmes bayésiens[74]. Rappelons brièvement les caractéristiques des trois classifieurs. Le système *Naïf-MARS* exploite uniquement les attributs de surface du système *Bio-MARS* et un classifieur bayésien naïf calcule le score du candidat. Le système *Naïf-CB* calcule aussi le score du candidat avec un classifieur bayésien naïf mais il dispose d'attributs linguistiques complexes

[74]Les résultats publiés dans les communications [Weissenbacher & Nazarenko, 2007a] et [Weissenbacher & Nazarenko, 2007b] sont erronés de 4 à 5% selon les systèmes. Les résultats que nous présentons ici ont été vérifiés manuellement.

Système	**Résultats**	
	Strict	*Partiel*
Aléatoire	6.42%	-
Premier GN	33.48%	42.97%
Bio-MARS	26.41%	37.05%
Naïf-MARS	40.32%	56.46%
Classifieur Bayésien Naïf	40.48%	56.57%
Classifieur Bayésien	39.60%	56.31%
MAX	63.71%	87.15%

TAB. 6.11 – Résultats de la résolution des anaphores pronominales sur le corpus Transcript (taux de succès)

Causes des erreurs	pourcentage des erreurs du système *Bio-MARS* sur la $1^{ère}$ itération
Sur-évaluation des scores des termes et des patrons de collocations	50%
Mauvais critères d'arbritage entre candidats égaux	38%
Erreur de prétraitement	12%
Total	100%

TAB. 6.12 – Typologie des erreurs du système *Bio-MARS* sur le corpus Transcript

en plus des attributs de surface du système *Naïf-MARS*. Le système *CB* exploite les attributs de surface et les attributs linguistiques complexes mais il calcule le score du candidat avec un RB ce qui lui offre mécanisme d'inférence pour atténuer les erreurs de calcul des valeurs de ces attributs. On s'attendait à ce que le système *CB* obtienne les meilleurs scores. Mais si on ne considère que le taux de succès strict il est moins performant que le système *Naïf-MARS*. Pourtant ses attributs, même partiels, semblent être un apport pour la classification puisque le système *Naïf-CB* est le meilleur système de notre expérience.

Pour rendre compte de ce classement nous avançons trois hypothèses concourantes. Le centre du contexte est le candidat qui a le plus de chance d'être l'antécédent mais son calcul s'avère difficile. Une heuristique consiste à rechercher le candidat saillant dans le contexte de l'anaphore puis à supposer que le candidat saillant et le centre du contexte sont identique. Notre hypothèse de travail était de calculer la saillance d'un candidat en employant toutes les informations à notre disposition, quelles que soient leurs fiabilités. Le mécanisme d'inférence du RB devait corriger l'incertitude d'un attribut grâce aux attributs qui le conditionnent.

Notre première hypothèse permet d'expliquer la proximité entre les résultats des systèmes : nous avons intégré un nombre d'attributs encore insuffisant pour améliorer significativement la classification du candidat central. L'étude des erreurs du système *CB* vient étayer cette hypothèse. Selon notre analyse, 70% des erreurs du système ont pour cause un mauvais calcul de l'élément central : 45% en raison d'un échec de la reconnaissance candidat saillant et 25%

en raison d'une disparité entre le candidat saillant et le candidat central. Ces erreurs dues au mauvais calcul de l'élément central auraient pu, en théorie, être évitées en exploitant d'autres connaissances.

La deuxième hypothèse est plus pessimiste : il existe un seuil de qualité pour les annotations syntaxiques et sémantiques en dessous duquel elles n'améliorent plus la résolution et la dégrade même un peu. Quelle que soit l'efficacité du mécanisme d'inférence, il ne peut trouver la bonne décision si les informations sur lesquelles il repose contiennent trop d'erreurs ou sont trop silencieuses. En comparant les erreurs des systèmes *Naïf-MARS* et *CB* nous avons remarqué qu'un certain nombre d'erreurs commises par le système *CB* proviennent d'une valeur erronée d'un attribut syntaxique ou sémantique. Nous avons alors renouvelé l'expérience en corrigeant et en complétant partiellement les annotations syntaxiques et sémantiques. Les nouveaux résultats montrent une nette amélioration en faveur du système *CB* et confirment que la qualité des annotations syntaxiques et sémantiques que nous avons utilisées est encore insuffisante pour la résolution anaphorique.

La dernière hypothèse s'ensuit immédiatement des performances des systèmes *CB* et *Naïf-CB* : la structure du RB qui différencie le système *CB* du *Naïf-CB* est trop simple. Elle n'exprime pas suffisament de contraintes et, en conséquence, ne corrige pas les valeurs erronées des attributs.

Nous développons chaque hypothèse et présentons les détails de nos résultats dans les sections suivantes.

Un nombre insuffisant d'attributs

Pour comprendre la raison d'une telle proximité entre les trois systèmes, nous avons procédé à l'analyse détaillée des erreurs du système qui aurait dû avoir les meilleurs résultats, le système *CB*. Nous avons focalisé notre analyse sur les erreurs de la première itération, dont nous donnons les caractéristiques dans le tableau 6.13.

Nombre total de pronoms	216
Nombre de pronoms impersonnels : 87	82 Vrais Positifs
	5 Faux Négatifs
Nombre de pronoms Anaphoriques : 129	126 Vrais Négatifs
	3 Faux Positifs
Anaphores résolues strictement	60
Anaphores résolues partiellement	16 (27 initialement avec 11 rejetées)

TAB. 6.13 – Résultat du système *CB* pour la $1^{ère}$ itération

Le tableau 6.14 résume les types d'erreurs du système que nous avons identifiés pour cette itération. Il apparaît que la majorité des erreurs proviennent d'un échec du système dans la reconnaissance de l'élément central. Soit parce que le candidat choisi par le système n'est pas le candidat saillant du contexte, soit parce que le candidat saillant qui est retrouvé correctement par le système diffère du candidat central du contexte. L'étude des erreurs montre que l'ajout d'informations linguistiques pourrait corriger une partie de ces erreurs en améliorant la discrimination du candidat saillant et exprimer des contraintes permettant de rejeter le candidat

lorsqu'il diffère du candidat central. Nous en donnons quelques exemples dans le paragraphe suivant et dans la section 6.3.3.

Causes des erreurs	Nombre d'erreurs	Pourcentage des erreurs du système *CB* sur la $1^{ère}$ itération
Calcul erroné du candidat saillant	18	27%
Anomalie du filtre du nombre	4	6%
Candidat saillant différent du candidat central	10	15%
Erreur de prétraitement	6	9%
Anaphore clausale	1	2%
Candidat partiellement retrouvé rejeté	11	17%
Candidat partiellement retrouvé accepté	16	24%
Total	66	100%

TAB. 6.14 – Typologie des erreurs du système *CB* sur le corpus Transcript lors de la $1^{ère}$ itération

Détail des erreurs du système *CB* : Lors de la première itération, à l'exception des candidats retrouvés partiellement (qu'ils soient acceptés ou rejetés), le système *CB* a produit 39 erreurs. 18 erreurs sont dues à un calcul erroné de l'élément saillant : le système ne retrouve pas ce que l'annotateur humain juge "intuitivement" être l'élément saillant parce qu'un nombre plus important d'attributs (les valeurs peuvent être correctes ou non) favorisent un candidat différent de l'élément saillant. Dans la phrase

[Gel filtration chromatography]$_1$ indicated that [the native enzyme]$_2$ existed as a dimer at high protein concentrations but that [it]$_2$ dissociated to a monomeric form on dilution.

l'antécédent est clairement l'élément saillant de la phrase et il obtient une bonne probabilité de 74%. Pourtant notre système lui préfère le candidat 1, lui ayant attaché une probabilité plus importante de 99%. Le système a correctement identifié que l'antécédent et le candidat sont tous deux sujets d'un verbe et qu'ils ont le même rôle grammatical que le pronom. L'antécédent est reconnu comme un GN défini et il suit un verbe spécifique. Le poids des ces attributs caractérisant l'antécédent aurait pu le qualifier. Mais les poids plus importants de la position du candidat comme premier GN de la proposition ainsi que la mauvaise annotation du mot *Gel* confondu avec le gène du même nom, lui donne l'avantage. Le fait que ce candidat soit indéfini et que le filtre de la cohérence sémantique le refuse ne suffit pas. Même si ce dernier attribut réduit la probabilité du candidat d'être l'antécédent, son poids est insuffisant pour l'éliminer, P(Coherence_Semantique=Incoherent|NonAntecedent)=0.0086 > P(Coherence_Semantique=Incoherent|Antecedent)=0.008. Cette dernière probabilité est, anormalement, non nulle en raison des valeurs erronées du corpus d'apprentissage. Notons, comme nous le verrons dans la section 6.3.3, que si nous corrigeons l'annotation du mot *Gel*, le système discrimine correctement le candidat saillant.

4 erreurs peuvent être expliquées par une anomalie du filtre du nombre. Cette anomalie provient d'une erreur de calcul de notre heuristique ou encore d'une mauvaise annotation morphologique et syntaxique. Prenons l'antécédent *one of the sites*, le GN *sites* est un pluriel correctement annoté et malheureusement filtré par notre heuristique qui rejette tout syntagme contenant

un mot au pluriel, de même pour l'antécédent *the ((G/U) AGCC) 11 RNAs* où le gène *RNAs* n'est pas reconnu par l'analyseur morphologique et étiqueté pluriel. Le cas de l'antécédent *the spo0B gene* est plus complexe. Nous nous reposons sur l'analyse syntaxique en constituants pour déterminer les GN du document. Cette analyse est bruitée. Elle concatène les deux syntagmes *the spo0B gene* et *genes* de la phrase :

Transcription of [the spo0B gene and genes] downstream of it was investigated by S1 nuclease protection experiments.

Notre système est donc contraint d'aligner l'antécédent *the spo0B gene* avec l'unique candidat *the spo0B gene and genes* qui est à son tour exclu par le filtre. Les rejets du filtre causés par une mauvaise analyse syntaxique des candidats sont les plus courants (3 erreurs sur 4).

Pour 10 erreurs, le système trouve bien le candidat qui paraît saillant à l'annotateur humain mais ce candidat diffère du centre qui est repris anaphoriquement. Prenons en exemple le contexte suivant :

The transcription of spoVE initiated within an hour after the onset of sporulation and coincided with the presence of RNA polymerase associated with a 33-kDa protein. [Amino acid sequence analysis]$_1$ of [the 33-kDa protein]$_2$ revealed that [it]$_2$ is a sigma factor, sigma E. Reconstitution analysis of sigma E purified from the sporulating cell extracts and vegetative core RNA polymerase showed that sigma E recognizes the P2 promoter.

Dans ce contexte l'auteur attire l'attention du lecteur sur les méthodes employées pour déterminer la nature de la protéine. Pour cela il positionne les deux analyses *Amino acid sequence analysis* et *Reconstitution analysis* en position de sujet. Mais, après la première phrase, le centre du discours se positionne sur la protéine *sigma E*. Notre système, qui assimile le centre à l'élément saillant ne choisit pas l'antécédent. En modélisant des connaissances propres du domaine, nous aurions pu intégrer au réseau une contrainte qui est violée par le candidat, un facteur sigma est une protéine, et rejeter ce dernier. Ce type d'erreur survient essentiellement lorsque l'antécédent est en position de complément du nom (avec 7 erreurs sur 10).

Les 7 erreurs restantes trouvent une explication simple. 4 antécédents sont absents de la liste des candidats. 2 pronoms anaphoriques mal résolus apparaissent dans une phrase dont la segmentation n'a pas été corrigée. Et enfin, 1 pronom réfère à un fait plutôt qu'à un objet du discours[75].

Une fiabilité trop faible des attributs linguistiques complexes

Au vu de l'avantage du système *Naïf-MARS* sur le système *CB*, nous devons nous interroger : la qualité des attributs linguistiques complexes calculés automatiquement sur ce corpus est-elle encore trop faible pour être un véritable apport au système, et ce, en dépit d'un mécanisme d'inférence sur des données incertaines ?

Pour nous en assurer nous avons mis en regard les erreurs du système *Naïf-MARS* avec celles du système *CB* sur la première itération. Le nombre de pronoms et l'exactitude du filtre de pronoms impersonnels sont donc identiques à ceux décrit dans le tableau 6.13, seul change le taux de succès du système de résolution. Sur cette itération le système Naïf-MARS ne retrouve strictement que 60 antécédents et 30 partiellement dont 13 ont été rejetés, le tableau 6.15 résume les chiffres. Afin de faciliter la comparaison nous les reportons dans le tableau 6.16.

[75] Nommément : *It is of interest that [the transcription of phage M2 DNA by both enzymes stringently requires*

Nombre total de pronoms	216
Nombre de pronoms impersonnels	82 Vrais Positifs
	5 Faux Négatifs
Nombre de pronoms Anaphoriques	126 Vrais Négatifs
	3 Faux Positifs
Anphores résolues strictement	60
Anphores résolues partiellement	17 (30 initialement avec 13 rejetées)

TAB. 6.15 – Résultat du système *Naïf-MARS* pour la $1^{ère}$ itération

Détail des erreurs des systèmes CB et Naïf-MARS	
16 candidats retrouvés partiellement par les 2 systèmes et acceptés	
11 candidats retrouvés partiellement par les 2 systèmes et refusés	
32 erreurs communes aux 2 systèmes	
4 antécédents manqués par *Naïf-MARS* et strictement retrouvés par le *CB*	4 antécédents manqués par le *CB* et strictement retrouvés par *Naïf-MARS*
	3 antécédents manqués par le *CB* et partiellement retrouvés par *Naïf-MARS* mais 2 sont rejetés
Naïf-MARS : 49 erreurs	CB : 50 erreurs

TAB. 6.16 – Détail des erreurs communes et propres aux systèmes *Naïf-MARS* et *CB*

Un rôle négatif des attributs linguistiques imparfaits : L'examen des erreurs du système *CB* ainsi qu'une estimation de l'imperfection des attributs linguistiques clé, que sont la classe sémantique et le rôle grammatical d'un syntagme, nous poussent à une réponse positive. Sur les 5 erreurs propres au système *CB*[76], 4 sont imputables aux valeurs inexactes de la classe sémantique et du rôle grammatical du candidat choisi par le système. Le calcul du rôle grammatical est réalisé avec une bonne précision, 84% des relations présentes sont correctes, mais un très mauvais rappel, seulement 12% des antécédents et 8% des candidats ont un rôle associé. A l'inverse TagEN recouvre 71% des noms de gènes et de protéines du corpus mais seulement 68% des entités étiquetées sont correctes (une baisse attribuable aux noms de gènes ou de protéines ambigüs tel que *not, All, Similar, RNA...*)

Un rôle positif des attributs linguistiques partiellement corrigés : Afin de nous assurer de l'impact du bruit sur les performances du système *CB*, nous avons prolongé notre expérience en modifiant certains paramètres. Nous avons corrigé les annotations relatives aux 47 pronoms résolus incorrectement par un des systèmes *CB* ou *Naïf-MARS* puis renouvelé la résolution avec les paramètres appris sur la totalité du corpus afin de comparer les nouveaux taux de succès. Nous avons "corrigé" manuellement les entités nommées, l'analyse syntaxique en constituants

KC1]$_1$. *[It]*$_1$ *may be due to this ion dependence that RPase L2 has not been detected previously.*

[76] C'est-à-dire les 4 antécédents manqués par le *CB* et strictement retrouvés par le *Naïf-MARS* plus l'antécédent manqué par le *CB* et partiellement retrouvé par *Naïf-MARS.*

de la phrase contenant l'antécédent et nous avons vérifié ou rajouté les rôles grammaticaux du pronom, de l'antécédent et du candidat préféré à l'antécédent. En considérant l'intégralité du corpus pour l'apprentissage des paramètres des classifieurs, nous introduisons sciemment un biais. Il a pour but de réduire l'effet des attributs bruités sur les paramètres en augmentant le nombre de données d'apprentissage. Les paramètres ainsi obtenus sont les meilleurs possibles pour notre corpus compte tenu de la qualité de son annotation. Le tableau 6.17 nous confirme qu'une fois les valeurs des attributs linguistiques corrigés, même partiellement[77], le système *CB* les exploite correctement et voit son taux de succès se détacher significativement de celui du système plus pauvre.

Détail des erreurs des systèmes CB et Naïf-MARS après une correction partielle	
16 erreurs communes aux 2 systèmes	
6 antécédents manqués par *Naïf-MARS* et strictement retrouvés par le *CB*	
6 antécédents manqués par *Naïf-MARS* et partiellement retrouvés par le *CB*, tous acceptés	5 antécédents manqués par le *CB* et partiellement retrouvés par *Naïf-MARS* mais 3 rejetés
Naïf-MARS : 28 erreurs strictes (16 erreurs partielles)	CB : 21 erreurs strites (10 erreurs partielles)
Naïf-MARS, nouveau taux de succès Strict :46.51%, Partiel :63.57%	CB, nouveau taux de succès Strict :52.71%, Partiel :73.64%

TAB. 6.17 – Détail des erreurs communes et propres aux systèmes *Naïf-MARS* et *CB* après une correction partielle des attributs linguistiques

Etudions cette amélioration sur notre précédent exemple. Lors de la première itération les deux systèmes choisissent, comme antécédent, le premier candidat de la phrase :

[Gel filtration chromatography]$_1$ indicated that [the native enzyme]$_2$ existed as a dimer at high protein concentrations but that [it]$_2$ dissociated to a monomeric form on dilution.

Le système *Naïf-MARS* sélectionne le candidat avec une forte fiabilité 82% car il est le premier GN de la phrase. Les poids des attributs décrivant l'antécédent, un GN défini et précédé d'un verbe spécifique, ne suffisent pas pour le qualifier (le système lui associe une fiabilité de 20% seulement). Le système *CB* possède plus d'attributs pour décrire les candidats concurrents, peut-être trop... Le système a à sa disposition les rôles grammaticaux des deux candidats et du pronom et il établit correctement le parallélisme syntaxique des GN : ils sont tous sujets d'un verbe. Mais le système, pour cet exemple, exploite aussi la classe sémantique des candidats. Une classe sémantique qui est erronée pour le premier candidat dont le mot *Gel* est confondu avec le gène du même nom, mais dont le poids est suffisamment important pour désigner le candidat aux dépens de l'antécédent (avec des scores respectifs de 98% et 74%). Dès lors que l'on a corrigé l'attribut erroné, le *CB* reconsidère correctement la position de l'antécédent qui, cette fois-ci, l'emporte sur le premier candidat grâce aux attributs du verbe spécifique et de l'article défini (avec un score de 97% pour le candidat contre 98% pour l'antécédent). Le système *Naïf-MARS* ne considérant pas les rôles grammaticaux des candidats voit ses entrées inchangées

[77] Notons qu'aucune correction n'a été apportée aux autres attributs comme les filtres ou les patrons de collocations.

(sinon le poids de l'attribut du premier GN augmenté par le nouveau paramétrage) et il infère de nouveau que *Gel filtration chromatography* est l'antécédent avec une probabilité de 84%.

Un mécanisme de renforcement inefficace

Les conclusions de la dernière expérience montrent qu'une correction partielle des données d'entrées suffit pour améliorer le score du *CB*. Ce rôle correctif est normalement dévolu au mécanisme de renforcement du système. Or ce mécanisme ne semble pas, ou peu, modérer le bruit des données. La confrontation des scores du *CB* et du *Naïf-CB* nous le confirme. Nous attendions des performances du *Naïf-CB* inférieures à celles du *CB*. Or leurs taux de succès sont très proches. La structure du *CB* apparaît, rétrospectivement, trop simple. Le nombre de contraintes exprimées par les relations entre les variables aléatoires ainsi que les valeurs de celles-ci sont insuffisantes, ce qui dégrade le processus de renforcement qui permet la correction du bruit des valeurs des attributs.

Envisageons, par exemple, le cas des erreurs du filtre du nombre. L'heuristique que nous utilisons pour déterminer le nombre d'un candidat peut produire des erreurs. Rappelons l'exemple de la section 6.3.3 où l'antécédent *the ((G/U) AGCC) 11 RNAs* est étiqueté pluriel. Lorsque cet exemple apparaît dans le corpus d'apprentissage nous apprenons une mauvaise probabilité : *P(Filtre_Nombre=Pluriel|Candidat=Antecedent)>0*, ce qui autorise le système à accepter des candidats pluriels lors de l'inférence. Sur l'ensemble des itérations du *CB*, 5 pronoms *we* n'ont pas été filtrés et ont été choisis comme antécédent par le système en raison de leur position dans la phrase. Or le système connaît par l'étiquetage morpho-syntaxique que ce candidat est un pronom personnel pluriel. Une information qu'il pourrait exploiter avec l'ajout d'une valeur mentionnant la nature d'un pronom pluriel dans la variable *SN_Pronom*. La nouvelle contrainte exprimée par l'ajout d'un arc entre les nœuds *SN_Pronom* et *Filtre_Nombre* interdirait à l'avenir le choix d'un pronom personnel pluriel comme antécédent.

6.4 Conclusion

Dans le chapitre précédent nous avons validé notre approche sur l'étape de filtrage des pronoms anaphoriques. Dans ce chapitre notre intention était de confirmer nos résultats sur un problème plus difficile, le choix de l'antécédent. Comme le problème de la distinction des pronoms impersonnels, le choix de l'antécédent se traduit aisément en une tâche de classification où l'opposition entre les systèmes exploitant des connaissances linguistiques complexes et les systèmes reposant sur des indices de surface est, historiquement, fortement marquée, ce qui fait de ce problème un très bon candidat pour notre étude.

Pour concevoir notre classifieur nous nous sommes inspiré du système pauvre en connaissance de [Mitkov, 2002], le système *MARS*. Nous avons conservé la stratégie générale de ce système. Nous recherchons toujours l'élément saillant d'un contexte anaphorique que nous supposons être aussi l'élément central du contexte, l'élement qui a la plus forte probabilité d'être l'antécédent du pronom. Mais nous avons compliqué le calcul de cet élément. D'une part nous avons abandonné la représentation des scores du système initial au profit des probabilités conditionnelles. Elles offrent une mesure continue plus fine du poids des indices et peuvent être apprises sur un corpus d'entraînement pour s'adapter au domaine du corpus. D'autre part,

conformément à notre approche, nous avons intégré au sein d'un même réseau bayésien les indices de surface du système initial et les connaissances linguistiques complexes associées. [Mitkov, 2002] a substitué aux connaissances linguistiques nécessaires pour trouver l'élément saillant, des indices de surface moins précis mais plus faciles et plus fiables à calculer. Notre réseau doit corroborer les connaissances linguistiques grâce aux indices de surface et corriger l'imprécision des indices grâce au connaissances complexes.

Pour évaluer notre approche nous avons comparé les performances des algorithmes de 7 systèmes de résolution. Les annotations d'entrée des algorithmes ont été produites par la même plateforme, la plateforme Ogmios, afin de ne pas introduire de biais dû à la différence de performance des outils d'annotation. Les systèmes sont identiques à l'exception des algorithmes employés pour choisir l'antécédent :

- Le système *Aléatoire* choisit l'antécédent aléatoirement.
- Le système *Premier GN* choisit toujours comme antécédent le premier GN de la phrase du pronom ou de la phrase précédente si le pronom est en début de phrase.
- Le système *MAX* choisit toujours l'antécédent si ce dernier apparaît dans la liste des candidats théoriques.
- Le système *Bio-MARS* choisit l'élément saillant du contexte comme antécédent ; l'élément saillant est calculé avec le système de score de l'algorithme original de [Mitkov, 1998] et la majorité des indices de surface employés dans la première version de l'algorithme.
- Le système *Naïf-MARS* choisit l'élément saillant du contexte comme antécédent mais l'élément saillant est déterminé par un classifieur bayésien naïf exploitant les indices de surface du système *Bio-MARS*.
- Le système *Naïf-CB* choisit l'élément saillant du contexte comme antécédent mais l'élément saillant est déterminé par un classifieur bayésien naïf exploitant les indices de surface du système *Bio-MARS* ainsi que les connaissances linguistiques complexes qui les complètent.
- Le système *CB* choisit l'élément saillant du contexte comme antécédent mais l'élément saillant est déterminé par un classifieur bayésien exploitant les indices de surface du système *Bio-MARS* ainsi que les connaissances linguistiques complexes qui les complètent.

Le système *Bio-MARS* est une version légèrement modifiée de l'algorithme original. Avant de l'utiliser comme système de référence, nous avons souhaité comparer les performances de notre système à celles du système d'origine, la première version du système *MARS*. Lors de notre première expérience, nous avons appliqué le système *Bio-MARS* sur le corpus d'évaluation du système *MARS* et comparé ses performances avec les performances publiées du système *MARS*. Durant cette expérience nous n'avons pas cherché à comparer strictement les résultats des systèmes en raison des différences entre les expériences. Le corpus d'évaluation que nous avons utilisé présente des différences avec le corpus utilisé lors de l'évaluation du système *MARS*. Les outils d'annotation qui composent les systèmes diffèrent et, en conséquence, la qualité des annotations d'entrée des algorithmes varie elle aussi. Enfin nous avons renoncé à certains indices que nous avons jugés trop spécialisés pour le corpus. En tenant compte de ces différences, l'écart de performance inférieur à 15% nous semble acceptable et nous autorise à utiliser notre implémentation comme système de référence lors de la deuxième expérience.

Pour évaluer notre approche nous avons réalisé une deuxième expérience sur deux corpus différents. Nous avons appliqué 6 systèmes sur le corpus MARS : les systèmes *Aléatoire*, *Pre-*

mier GN, *MAX*, *Bio-MARS* comme systèmes de références et les systèmes à évaluer *Naïf-MARS* et *CB*. Puis, nous avons conservé les systèmes de référence et évalué le système *Naïf-CB* en plus des systèmes *Naïf-MARS* et *CB* sur un corpus d'un autre domaine, le corpus Transcript. Lors de l'analyse des résultats nous avons été particulièrement attentif aux résolutions divergentes des systèmes. La mise en regard des résolutions contraires de deux systèmes nous a permis de comprendre les effets de chaque évolution que nous avons apportée au système initial *Bio-MARS*.

En comparant le comportement des systèmes *Bio-MARS* et *Naïf-MARS* sur un même corpus, nous vérifions l'intérêt de notre représentation probabiliste : une représentation plus fine de l'incertitude et un apprentissage du poids des attributs au corpus. Le système *Bio-MARS* obtient les meilleures performances sur le corpus MARS. Les scores et les heuristiques d'arbitrage entre deux candidats égaux du système *Bio-MARS* ont été déterminés humainement sur le corpus MARS et sont donc bien adaptés au corpus. Mais ce système est incapable de s'adapter à un nouveau corpus et il voit ses performances chuter significativement sur le corpus Transcript. Celles du système *Naïf-MARS* auraient dû être similaires à celles de *Bio-MARS* sur le corpus MARS mais elles sont décevantes. L'étape d'apprentissage des paramètres auraient dû adapter le poids des attributs au corpus mais la taille du corpus s'est révélée trop faible et les paramètres appris incorrects. Ses performances s'améliorent nettement sur le corpus Transcript dont la taille est plus importante. Les paramètres sont adaptés au corpus et la représentation continue de l'incertitude diminue fortement les cas d'égalité entre les candidats, limitant ainsi le recours aux heuristiques inadaptées pour ce corpus.

En comparant les erreurs respectives des systèmes *Naïf-MARS* et *CB* nous avons pu caractériser les conditions nécessaires pour la contribution des connaissances linguistiques complexes : ces connaissances améliorent la résolution si, sans être parfaites, elles sont de bonne qualité ; dans le cas contraire leur utilisation peut dégrader faiblement les performances du système. Le système *CB* obtient de meilleures performances que le système *Naïf-MARS* sur le corpus MARS alors que le rapport s'inverse légèrement sur le corpus Transcript. Les outils de notre plateforme de traitement étant spécialisés pour le domaine de la génomique, sur le corpus MARS le système *CB* ne dispose que des connaissances syntaxiques. Mais l'analyse syntaxique du corpus est d'une qualité suffisante pour améliorer la résolution du système *CB* dans certains cas. Pour traiter le corpus Transcript le système *CB* a, à disposition, des informations sémantiques en plus des informations syntaxiques. Mais la langue de spécialité du corpus est plus difficile à traiter et nos outils produisent une annotation de moins bonne qualité. L'analyse d'étaillée des résultats révèle que les attributs traduisant les connaissances linguistiques complexes, lorsqu'ils sont bruités, peuvent dans un petit nombre de cas tromper le système. Afin de confirmer notre résultat, nous avons corrigé certains attributs et renouvelé l'évaluation des deux systèmes : avec des connaissances linguistiques de meilleure qualité, bien que toujours imparfaites, les performances du système *CB* sont nettement supérieures à celles du système *Naïf-MARS*.

Le système *CB* est d'autant plus sensible au bruit des attributs d'entrée que le mécanisme de correction du système se révèle être peu efficace. La comparaison des performances *CB* et *Naïf-CB* nous le montre clairement. Le système *Naïf-CB*, qui suppose tous les attributs indépendants les uns des autres, réalise une résolution presque identique à celle du *CB*. La structure du réseau du système *CB* nous semble, après examen, trop simple : d'autres contraintes peuvent encore être exprimées et permettraient de corriger certaines erreurs du système.

Si les résultats de notre *CB* sur la tâche du choix de l'antécédent sont analogues aux résultats des systèmes les plus connus de l'état de l'art, ils sont inférieurs aux résultats que nous espérions

après la section précédente. Et l'écart entre ses performances maximales, obtenues lorsqu'il exploite des attributs en partie corrigés, et les performances du système *MAX* reste important (73.64% contre 87.15%). Toutefois, il s'agit là de la première version de notre système, nous pouvons encore l'améliorer en corrigeant les imperfections que nous avons constatées lors des premières expériences.

Chapitre 7

Prototype d'un réseau bayésien dynamique pour la résolution des anaphores

Le fait est, que les résultats principaux de la précédente section ne sont pas ceux que nous attendions. Pourtant, au terme des comparaisons entre les systèmes bayésiens, nous ne remettons pas en cause notre approche. Les performances de nos classifieurs sont comparables aux performances des systèmes de résolution d'anaphores de l'état de l'art. De plus, les résultats que nous avons obtenus restent partiels. Le mécanisme d'inférence proposé dans la section précédente s'est avéré relativement inefficace mais peut être amélioré simplement en modifiant le réseau. Ceci nous conduit, dans la section 7.4 de ce chapitre, à proposer un prototype reposant sur un réseau bayésien dynamique pour la résolution des anaphores.

7.1 Intérêt de l'approche probabiliste

Nous avons mis en évidence différents intérêts de l'approche probabiliste. Les scores bien meilleurs du système *Naïf-MARS*, comparés aux résultats du système *Bio-MARS*, ne sont dus qu'au changement de représentation et à l'apprentissage. Ceci confirme le résultat du chapitre 5.3.3 : un système probabiliste tolère dans une certaine mesure le changement de domaine de corpus en adaptant les poids des attributs utilisés dans la classification au nouveau domaine.

De plus, comme il a été noté ci-dessus, le passage aux probabilités a introduit une mesure continue plus précise que la mesure discrète du système *Bio-MARS*. Et il est intéressant de noter que lorsque le système commet une erreur stricte, l'antécédent est proposé en deuxième solution dans 49% des cas pour le système *Naïf-MARS* et dans 58% des cas pour le système *CB* sur la première itération. Les taux de succès des deux systèmes interdisent de se fier aveuglément à leurs résolutions automatiques dans une application de TAL, mais l'ajout d'une interface de validation manuelle des deux premiers candidats proposés augmenterait considérablement la fiabilité de la résolution sans épuiser le temps et l'énergie de l'annotateur. Les taux de succès stricts calculés sur les deux premières propositions des systèmes *Naïf-MARS* et *CB* sur la première itération sont respectivement de 59.69% et 64.34%. Une étude récente de [Mitkov *et al.*, 2007] a déjà mesuré une amélioration d'un système d'extraction d'informations, de résumés et d'une

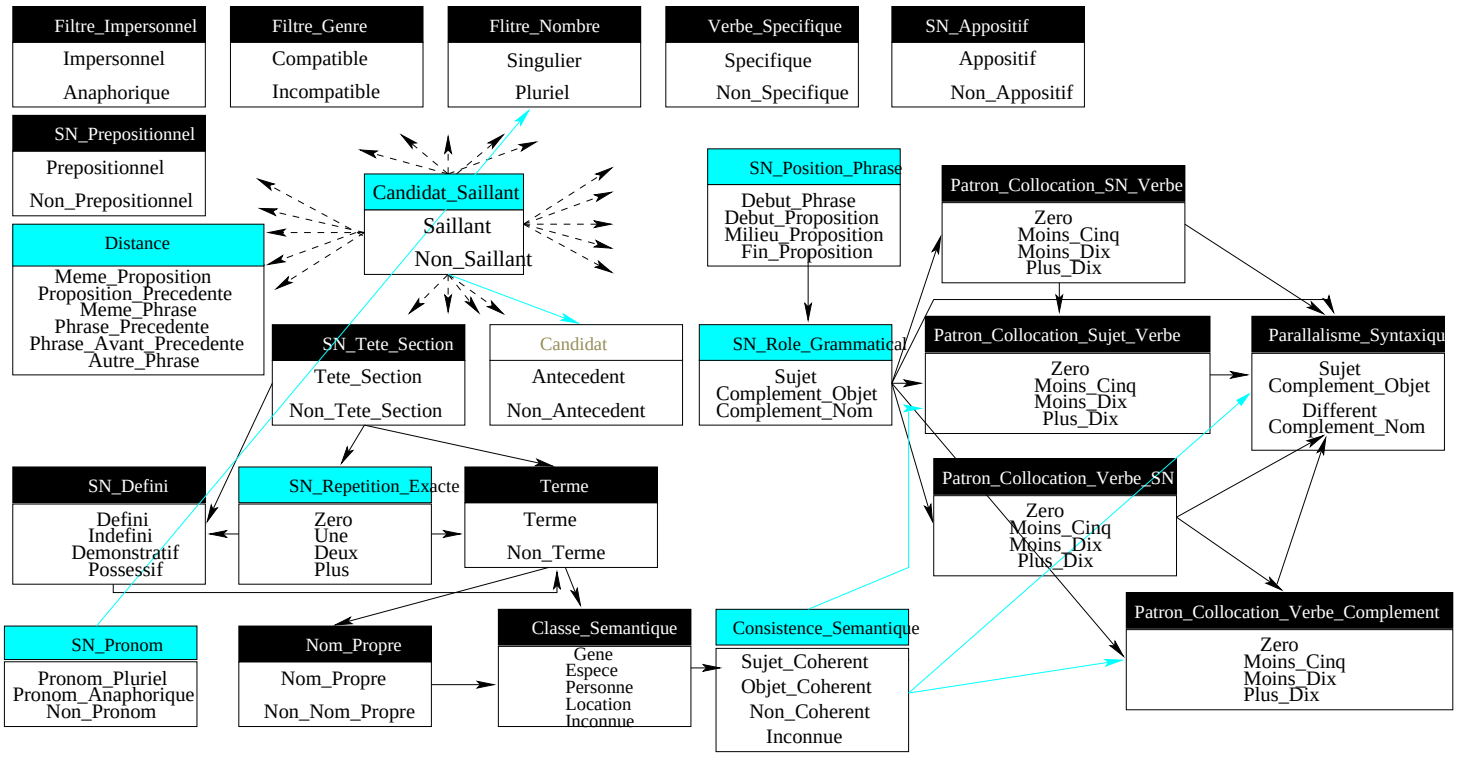

FIG. 7.1 – Corrections apportées au classifieur bayésien pour la résolution des anaphores

classification automatique de documents embarquant le système MARS (deuxième version). En proposant aux systèmes de TAL, le candidat proposé par une version semi-automatique de notre système validé humainement, nous sommes en droit d'attendre un effet plus important de la résolution des anaphores dans une tâche de TAL.

Enfin, l'efficacité du raisonnement probabiliste réalisé avec un réseau bayésien dépend de la complexité des VA et de la structure de ce dernier. Si l'intérêt de ce raisonnement a été mis en évidence dans le chapitre 5, il a été faiblement confirmé dans le chapitre précédent en raison de la modélisation d'un réseau trop simple. Nous proposons dans la section suivante une révision de la modélisation des réseaux utilisés lors de cette étude en vue d'améliorer leurs performances.

7.2 Améliorer le Classifieur Bayésien pour le choix du candidat

7.2.1 Corriger le réseau

Corriger la représentation de certains attributs

La première correction que nous souhaitons apporter au réseau est une représentation plus précise de la phrase et du nombre de répétitions du candidat. Les modifications du réseau sont distinguées par la couleur bleue (ou gris clair) dans la figure 7.1.

Nous autorisons une représentation plus fine de la phrase grâce aux nouvelles valeurs des nœuds `Distance` et `Premier_SN` qui a été renommé `Position_NP_In_Sentence`, un nom plus approprié. La nouvelle segmentation de la phrase précise la distance entre le candidat et le pronom avec une mesure réalisée au niveau de la proposition et non plus de la phrase. Cette mesure devrait améliorer la représentation de l'introduction d'un candidat dans le discours qui est une information essentielle dans le calcul de la saillance de ce dernier. La position du candidat calculée au niveau de la proposition nous permettera dans la section 7.2.3 de modifier la fiabilité du rôle grammatical du candidat.

Le nœud `Repetition_NP` a lui aussi été rebaptisé `SN_Repetition_Exacte` afin de le distinguer des nouveaux attributs que nous avons introduit dans la section 7.2.2, pour cacluler les répétitions des syntagmes synonymes, hyperonymes ou encore partageant la même tête. Le compte des répétitions a lui aussi été remanié, il est maintenant plus précis.

Dans la première version de notre réseau l'élément central était confondu avec l'élément saillant. Le nœud de décision `Candidat` évaluait donc la probabilité pour le candidat d'être l'élément saillant du contexte anaphorique. Nous avons distingué le nœud de décision `Candidat` du nœud `Candidat_Saillant` pour permettre un calcul dédié à l'élément central et le différencier de l'élément saillant, nous discutons ce point dans la section 7.2.2. Le nœud `Candidat_Saillant` est relié à l'ensemble des attributs employés pour calculer la saillance du candidat mais il perd son statut de nœud de décision.

Enfin la VA *SN_Sujet* est remplacée par la VA *SN_Role_Grammatical*. Les valeurs de la VA *SN_Sujet* sont inadaptées à notre problème. Nous avons vu que l'analyse syntaxique en dépendance est très silencieuse. La valeur *Inconnu*, initialement employée pour faciliter l'apprentissage des paramètres, contribue à la perte d'information de l'attribut. Les données étiquetées *Sujet* ou *Complément* sont trop peu nombreuses dans notre corpus d'apprentissage et les probabilités *a priori* apprises pour cet attribut sont incorrectes. Lors de la phase d'inférence, par exemple, l'ajout de l'observation *First* augmente seulement de 0.2% la probabilité *a priori* pour le candidat d'être le sujet, une augmentation qui est bien trop faible en comparaison de celle logiquement attendue par l'expert. La valeur *Inconnu* doit être supprimée au profit d'un détail plus grand dans les rôles grammaticaux du candidat. Nous pensons notamment à la distinction des rôles complément d'objet et complément du nom[78]. Bien sûr, ces modifications doivent être accompagnées d'un apprentissage des paramètres plus complexe : nous devrons utiliser un algorithme d'apprentissage capable d'estimer les valeurs manquantes des données d'entraînement.

Renforcer les connaissances du réseau

Dans la section 6.3.3 nous avons déjà suggéré l'ajout de la valeur *Pronom_Pluriel* pour le nœud `SN_Pronom` et de l'arc entre ce nœud et le nœud `Filtre_Nombre`. Cette contrainte permet de corriger certaines erreurs du filtre du nombre grâce aux étiquettes morphosyntaxiques du candidat.

De la même manière, nous avons ajouté des arcs entre le nœud `Semantic_Consistence` et les nœuds `Patron_Collocation_Sujet_Verbe`,

[78] Afin de garder la cohérence du réseau, nous modifions aussi les valeurs du nœud `Consistance_Semantique` (voir la figure 7.1).

`Collocation_Verb_Complement_Pattern` et `Parallelisme_Syntaxique`, dans le but de renforcer les connaissances syntaxiques grâce aux connaissances sémantiques. Pour comprendre ce renforcement, considérons une nouvelle fois l'exemple de la section 6.1.2. Dans cet exemple le pronom anaphorique est le sujet du verbe *encode*. Parmi les candidats possibles à l'antécédence, nous retrouvons le candidat *another ORF*. Supposons que le système ne retrouve aucune collocation où *ORF* est le sujet du verbe *encode*, mais qu'il sait que le candidat *ORF* est un gène. Il peut déterminer la cohérence sémantique entre ce candidat et notre pronom en recherchant les collocations où un gène est le sujet du verbe *encode* : <GENE/encode>. Or de telles collocations existent dans notre corpus[79], nous pouvons donc logiquement conclure que le candidat *ORF* aurait pu être le sujet du verbe *encode* et que la collocation <ORF/encode> est possible. L'arc que nous avons ajouté entre les nœuds `Semantic_Consistence` et `Patron_Collocation_Sujet_Verbe` permet de renforcer l'existence d'une telle collocation à partir de la cohérence sémantique. Cette collocation possible renforce aussi le parallélisme syntaxique : si un gène peut être le sujet du verbe *encode* et que le candidat *ORF* est un gène, nous pouvons en déduire qu'il peut, comme le pronom, être aussi le sujet du verbe *encode*.

Apprendre les paramètres sur des données corrigées

Durant notre précédente expérience le système a produit une série de données relatives aux antécédents et aux autres candidats de notre corpus. Ces données peuvent maintenant être complétées et corrigées en vue d'être intégrées comme des données d'expert pour une nouvelle résolution sur un corpus de génomique différent. Intégrées aux données du nouveau corpus d'apprentissage calculées automatiquement, ces données de meilleure qualité garantiront une certaine fiabilité pour les résultats d'un apprentissage plus complexe sur cet ensemble de données qui seront encore bruitées et manquantes.

7.2.2 Enrichir le réseau de nouvelles connaissances

Les attributs linguistiques complexes qui ont complété le système Bio-MARS étaient peu nombreux. Les connaissances sémantiques nécessaires à la résolution des anaphores étaient quasiment absentes de notre système et si bruitées que l'apport fut très faible. Or des connaissances syntaxiques, sémantiques ou ontologiques plus complètes, qu'elles soient générales ou spécialisées, existent et peuvent enrichir notre système. Nous ne dessinons ici qu'un prototype encore incomplet d'un réseau bayésien, pour la résolution des anaphores, spécialisé pour les textes de génomique. Une étude des connaissances supplémentaires et d'un corpus de développement plus important que celui utilisé pour les expériences précédentes est nécessaire pour imaginer des liens pertinents oubliés, corriger leur orientation, confirmer la bonne "granularité" des attributs et concevoir le meilleur algorithme de calcul des valeurs de ces derniers. Le réseau complet étant trop chargé pour être lisible, nous commenterons les sous-graphes des modifications.

[79] Rappelons par exemple la phrase :
The [sacT]$_{gene}$ gene which controls the sacPA operon of Bacillus subtilis [encodes] a polypeptide homologous to the B. subtilis SacY and the Escherichia coli BglG antiterminators.

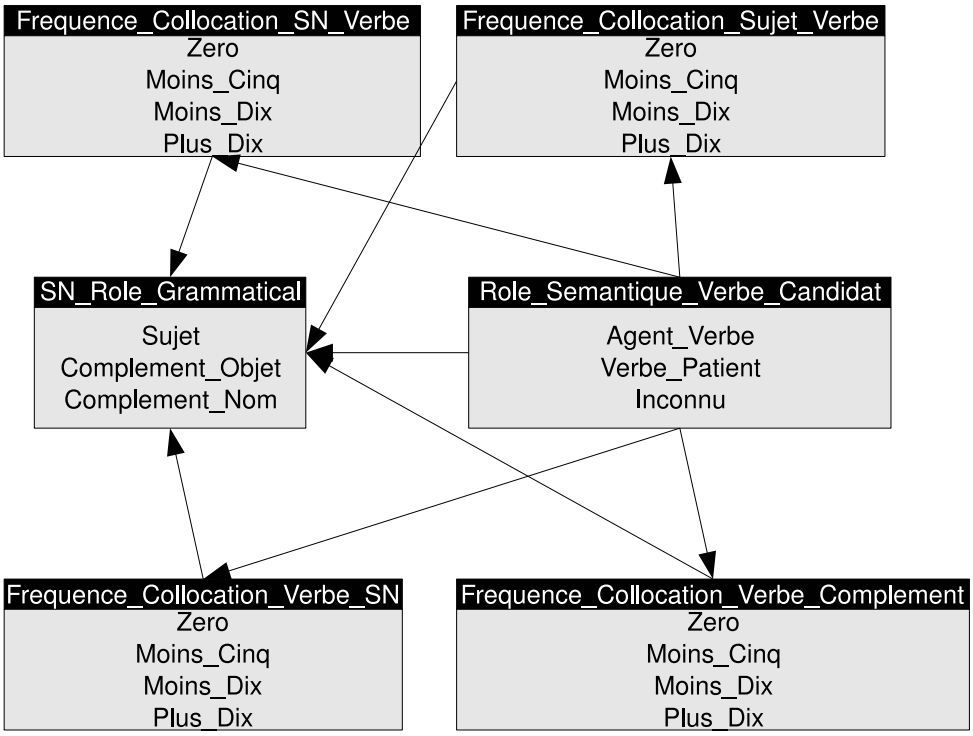

FIG. 7.2 – Attributs renforçant le rôle grammatical du candidat

Des connaissances sémantiques pour renforcer les connaissances syntaxiques

Une ressource générale, les rôles sémantiques, a récemment été étudiée et intégrée dans le processus de résolution des anaphores [Ponzetto & Strube, 2006]. Cette ressource sémantique peut être utilisée de deux manières différentes pour renforcer différentes connaissances du réseau. Considérons le contexte anaphorique général suivant pour rendre la présentation plus claire :

Candidat verbe$_1$. *Pronom verbe*$_2$. où le candidat est le sujet du *verbe*$_1$ et le pronom du *verbe*$_2$

Les rôles sémantiques peuvent être utilisés pour renforcer les probabilités de différents attributs syntaxiques comme le montre la figure 7.2. Dans cette figure nous avons introduit 4 nouveaux nœuds, en plus du nœud `SN_Role_Grammatical`, qui sont tous influencés directement par la connaissance du rôle sémantique du candidat :

- l'attribut `Frequence_Collocation_SN_Verbe` (respectivement `Frequence_Collocation_Verbe_SN`) représente le nombre de phrase dans notre corpus où le *verbe*$_1$ est le premier verbe à suivre (resp. à précéder) le candidat,
- l'attribut `Frequence_Collocation_Sujet_Verbe` (respectivement `Frequence_Collocation_Verbe_Complement`) représente le nombre de phrase dans notre corpus où le candidat a été retrouvé sujet (resp. complément) du *verbe*$_1$.

L'attribut `Role_Semantique_Verbe_Candidat` permet d'intégrer les rôles sémantiques au réseau et de renforcer les probabilités des nœuds précédents : lorsque la tête du can-

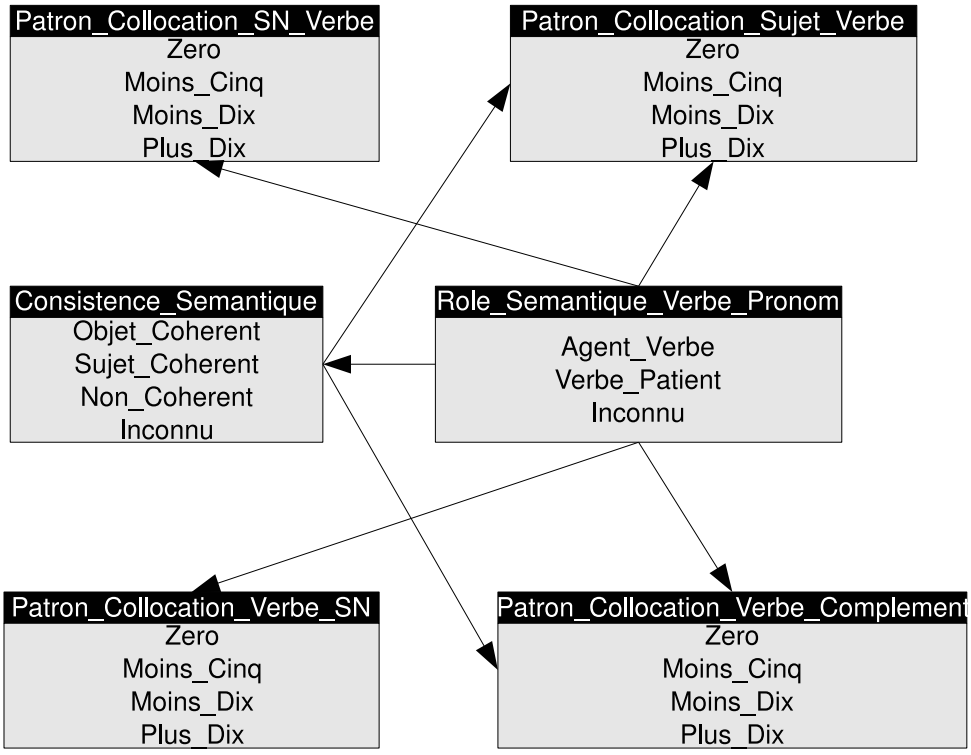

FIG. 7.3 – Attributs renforçant les patrons de collocations

didat est un agent connu du $verbe_1$, que l'analyseur ait trouvé ou non cette relation, il est très probable que le candidat soit le sujet du $verbe_1$, ce qui renforce la probabilité de trouver des collocations <*sujet-verbe*$_1$> ou <*Candidat/Verbe*$_1$> et diminue la probabilité des autres collocations (et inversement lorsque le candidat est un patient possible du $verbe_1$).

De manière analogue, les rôles sémantiques peuvent aussi être employés pour renforcer les patrons de collocations et la consistence sémantique entre le pronom et le candidat, comme le suggère la figure 7.3. Si, par exemple, la tête du candidat est un agent connu du verbe dont le pronom est le sujet, *i.e.* du $verbe_2$, le candidat est un sujet possible du $verbe_2$ (ce qui renforce la probabilité de trouver des patrons collocations <*sujet-verbe*$_2$> ou <*Candidat/Verbe*$_2$>) et il est cohérent sémantiquement avec le pronom.

Des connaissances sémantiques pour un calcul plus complet de la répétition

Les classes sémantiques de *WordNet* et leurs hiérarchies pour le calcul de différents attributs comme la tête du syntagme, la cohérence sémantique ou encore la répétition des GN synonymes, ont déjà été utilisées avec succès (voir [Mitkov, 2002]). Le recours à des ressources spécifiques au domaine pourrait suppléér complètement ou partiellement les ressources générales comme *WordNet*. Pour le domaine de la génomique de nombreuses ressources terminologiques et ontologiques existent et s'enrichissent quotidiennement. *Wordnet* pourrait être avantageusement remplacé par *Gene Ontology*[80] par exemple.

[80] http://www.geneontology.org/

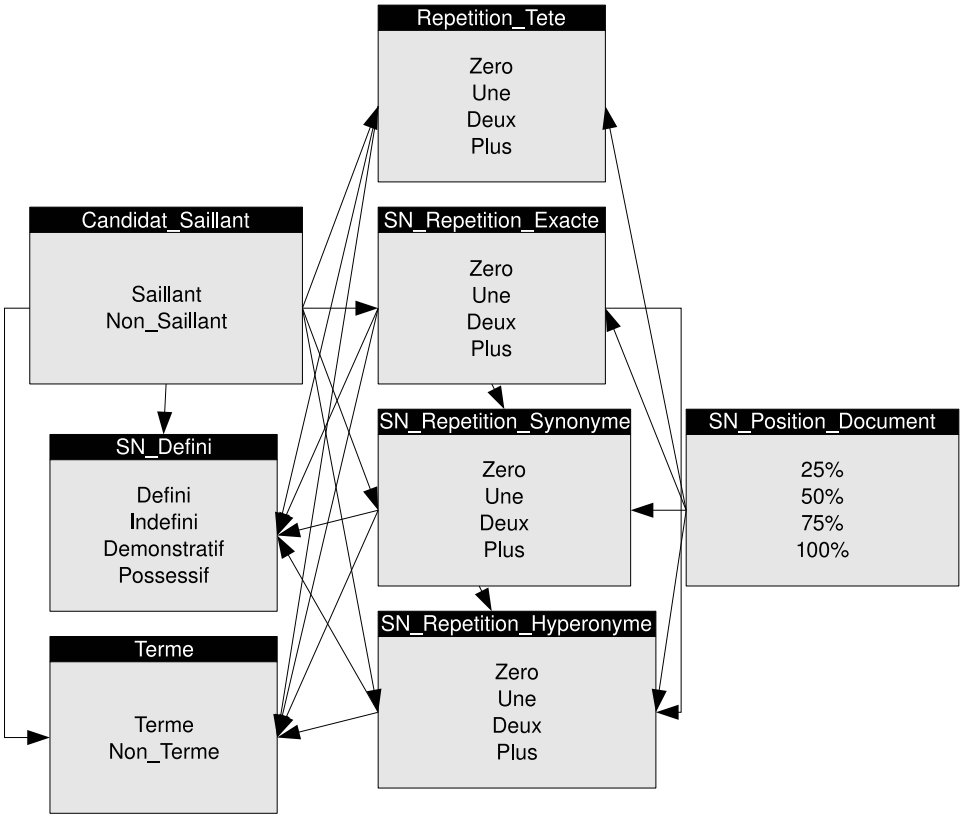

FIG. 7.4 – Attributs de répétition d'un candidat

Un emploi immédiat de ces ressources pour notre réseau serait un meilleur calcul du facteur de répétition d'un candidat donné. Nous proposons dans la figure 7.4 quatre attributs permettant de traquer la répétition. Les attributs `Repetition_Tete`, `SN_Repetition_Exacte` et `SN_Repetition_Synonyme` sont directement empruntés à la deuxième version du système MARS. Lorsqu'un candidat est répété à plusieurs reprises dans le discours il est plus saillant. Nous avons calculé les répétitions de la manière la plus simple : un candidat est répété lorsque nous retrouvons ce candidat à l'identique dans la suite du document. Deux autres façons de calculer ce facteur seraient de considérer tous les synonymes et les hyperonymes immédiats d'un candidat présents dans la suite du document comme une réptition de celui-ci. Dans l'exemple suivant :

> *AML1 is also known as CBFA2. This gene is one of the most frequently translocated or mutated genes in human cancer (7-12).*

le candidat *AML1* est répété deux fois, une première fois avec le synonyme *CBFA2* et une seconde fois avec son hyperonyme *This gene*. Une dernière manière de calculer la répétition d'un candidat consiste à ajouter au nombre de répétitions les candidats qui partagent la même tête que le candidat même si ces candidats désignent un autre objet du discours comme les candidats *the first channel* et *the second channel*.

Le calcul des répétitions de l'hyperonyme immédiat et des synomymes d'un candidat nécessite une ontologie pour être calculée. Son impact en tant que facteur de saillance doit encore être soumis à une étude sur un corpus de développement. De même nous souhaitons étudier l'impact de la position du candidat dans le document grâce à l'attribut `Position_NP_In_Document`. Un candidat qui apparaît en début de document à peu de chance d'être répété alors qu'il en a beaucoup plus en fin de document. Nous supposons donc que les poids des attributs de répétition dans la saillance doivent être dépendants de la position du candidat dans le document et nous les mesurons avec les arcs allant du nœud `Position_NP_In_Document` aux nœuds des attributs de répétitions.

Des connaissances syntaxiques pour renforcer la saillance et le rôle grammatical

L'omission des indices linguistiques tels que les indices de *Référence immédiate* ou *Instruction séquentielle* s'est révélée être une mauvaise décision. Ces connaissances syntaxiques qui peuvent se comprendre comme des règles logiques permettent, en décrivant finement des structures linguistiques particulières, de focaliser avec fiabilité le système sur l'élément saillant. Mais l'utilisation de ces indices reste incommode puisqu'il faut réécrire l'ensemble des règles selon le domaine et le genre du corpus. Pour notre corpus, la présence de ce type de règles aurait certainement prévenu certaines erreurs partielles. Un ensemble d'expressions régulières pourrait reconnaître les constructions où le complément du nom est l'antécédent du pronom de la proposition suivante. Considérons l'expression :

– `GN_ANALYSE` *of* `GN_antecedent` `Verbe_REVELER` *that it* `VERBE`

où `GN_ANALYSE` est un groupe nominal dont la tête appartient à une classe sémantique marquant l'étude d'un objet génétique, `GN_antecedent` est l'antécédent du pronom qui suit, `Verbe_REVELER` est un verbe d'une classe sémantique regroupant des synonymes de *reveal* ou *suggest* et enfin *VERBE* un verbe différent des verbes de la classe `Verbe_REVELER`.

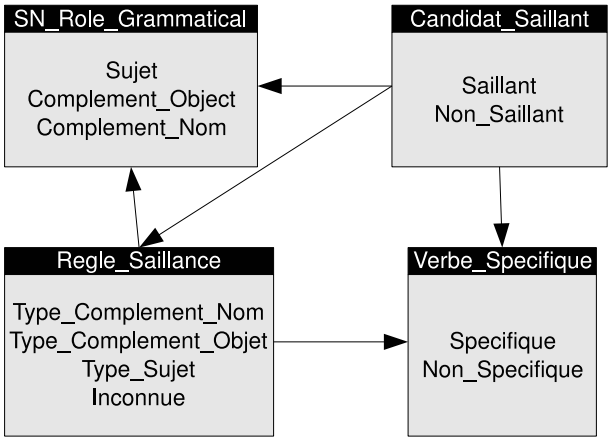

FIG. 7.5 – Attributs dénotant un candidat saillant et un rôle grammatical défini

Cette règle corrigerait la mauvaise résolution de notre système qui choisit le GN entier d'indice 1 des phrases :

- *[An SDS-polyacrylamide gel analysis of [the most purified preparation]$_2$]$_1$ revealed that [it]$_2$ consists of beta, beta', alpha, and sigma with apparent molecular masses of 151, 147, 42, and 55 kDa, respectively.*
- *[Biochemical characterization of [the protein]$_2$]$_1$ revealed that [it]$_2$ is an FMN-containing NADPH oxidase.*
- *[Analysis of the amino-terminal residues of [the protein produced in Escherichia coli]$_2$]$_1$ suggests that [it]$_2$ is processed by a methionine aminopeptidase.*

Ce type de règles a la propriété d'indiquer avec une bonne certitude que le GN que l'on suppose être l'antécédent du pronom est le complément du nom du GN `GN_ANALYSE`. Les règles logiques de Mitkov ont la même propriété puisqu'elles indiquent clairement les compléments d'objets du verbe de la séquence reconnue. Sur la figure 7.5 un arc a été introduit entre le nœud `Salient_Rule` et `Verbe_Specifique` car les règles de saillance jouent un rôle identique à celui des verbes spécifiques et pourraient, selon les règles, renforcer le poids d'un GN succédant à un verbe spécifique.

Une recherche dédiée au candidat central

Jusqu'à présent, notre résolution a reposé uniquement sur un calcul simple de la saillance en vue de se soustraire au calcul plus complexe du centre (voir la section 2.2.1). Cette stratégie est connue pour échouer sur des phrases longues et complexes [Mitkov, 2002], ce que nous avons pu constater à notre tour sur notre corpus : sur les 40 erreurs strictes du *CB*, seulement deux erreurs concernaient une anaphore interphrastique[81].

Avec l'emploi des connaissances sémantiques et ontologiques spécialisées nous sommes en mesure de déterminer le centre d'une proposition, ou, tout au moins, de construire une liste des éléments centraux d'un contexte anaphorique. Il est alors tout à fait possible d'intégrer

[81] La proportion d'anaphores interphrastiques et intraphrastiques à résoudre dans l'itération est comparable à celle du corpus général avec 70% d'anaphores intraphrastiques.

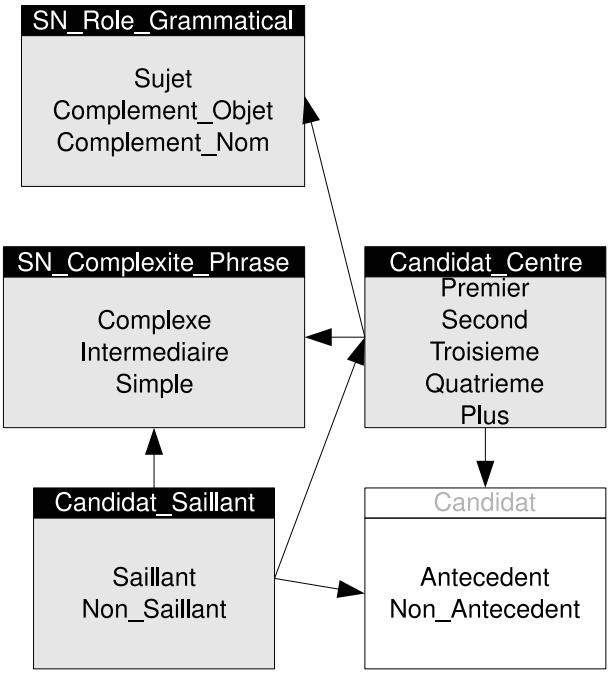

FIG. 7.6 – Ajout du centre dans du calcul de l'antécédent

cette nouvelle information complémentaire de la saillance au réseau. Le choix de l'algorithme employé pour traquer le centre reste tenu à un examen des nombreuses adaptations et implémentations issues de la théorie du centre. Cet examen sort du cadre de notre travail, toutefois le lecteur intéressé pourra trouver une étude préliminaire dans [Mitkov, 2002].

Dans la figure 7.6 nous avons intégré un nouvel attribut pour le calcul du centre, `Candidat_Centre`. Les valeurs de cet attribut déterminent la position du candidat dans la liste des éléments centraux d'un contexte anaphorique que nous calculons. Si notre candidat est le premier élement de la liste, il est l'élement central. Nous avons distingué l'élément central de l'élément saillant mais comme ils se confondent souvent lorsque les phrases sont simples, les deux attributs ne sont pas indépendants et un arc les relie. Lorsque la phrase est plus complexe, il peut être sage de dissocier le centre et l'élément saillant comme nous l'avons vu précédemment, une séparation réalisée grâce aux arcs vers le nœud `Complexity_NP_Sentence`, qui permettent de varier les poids de chaque attribut selon la complexité de la phrase. La décision est inférée par la probabilité du nœud `Candidat`. Beaucoup d'algorithmes de calcul du centre, dont [Brennan *et al.*, 1987], ordonnent la liste des candidats au moyen de leur rôle grammatical. Nous modérons alors le poids du centre par la probabilité du rôle grammatical supposé du candidat. En l'absence d'un algorithme de calcul précis nous n'ajouterons pas d'autre attribut ou corrélation du poids du centre.

7.2.3 Douter des observations

Jusqu'ici, nous n'avons jamais douté des valeurs des données d'entrée. Nous avons affecté chacune de nos observations d'une certitude maximale, ce qui est loin d'être raisonnable. Des

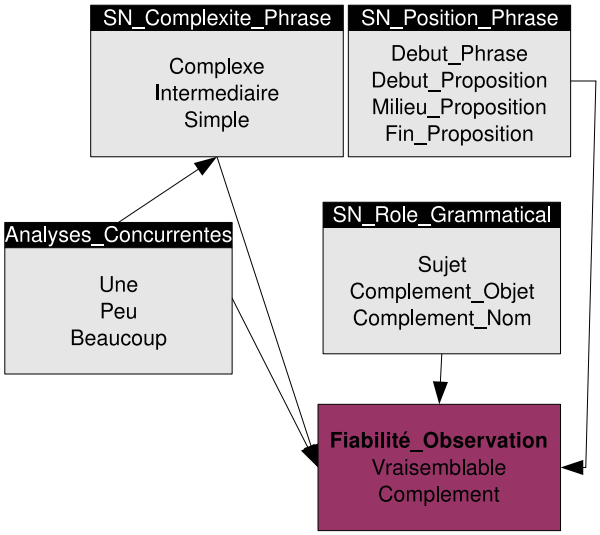

FIG. 7.7 – Attributs responsables de la variation de la fiabilité du rôle grammatical du candidat

vérifications élémentaires permettent d'affermir ou non notre certitude quant à la valeur d'une observation.

Douter du rôle grammatical

Une première vérification de la fiabilité de l'annotation peut être faite en interrogeant l'outil qui produit l'annotation. Certains outils déterminent un degré de confiance des annotations qu'ils ajoutent aux documents ou proposent plusieurs annotations concurrentes pour une même séquence. Pour des phrases complexes, le *Link Parser* peut proposer différentes analyses ordonnées de la meilleure à la moins bonne, la meilleure n'étant pas nécessairement l'analyse correcte. Dans une telle situation les rôles grammaticaux sont évidemment moins fiables. Une probabilité pourrait être estimée en comptant le nombre de répétitions d'un rôle au sein des différentes analyses. Une heuristique que nous n'avons pas pu implémenter jusqu'ici, la plateforme *Ogmios* ne conservant que la première analyse et ne mentionnant pas la fiabilité des annotations.

Dans la figure 7.7 le nœud `Fiabilite_Observation` bien que marqué d'une couleur différente des autres nœuds est d'une nature identique, mais nous le distinguons car son rôle est de faire évoluer la fiabilité du rôle grammatical. Cette fiabilité varie selon la complexité de la phrase, la position du candidat dans cette dernière et le nombre d'analyses concurrentes proposées par un ou plusieurs analyseurs syntaxiques. Plus la phrase est complexe, plus il est difficile pour l'analyseur de l'analyser entièrement et de donner une unique solution. L'observation du nombre d'analyses concurrentes est évidente et permet donc de renforcer l'estimation de la complexité de la phrase dont le calcul reste soumis à décision. Toutefois, si le candidat est situé au début de la phrase ou d'une proposition, l'analyseur, qui calcule le rôle grammatical de la tête du candidat localement, a plus de chances de le trouver correctement et ce, indépendamment de la complexité de la phrase ou de la proposition. La chance décroît à mesure que le candidat se situe en milieu et en fin de proposition. Nous le matérialisons par un arc entre les

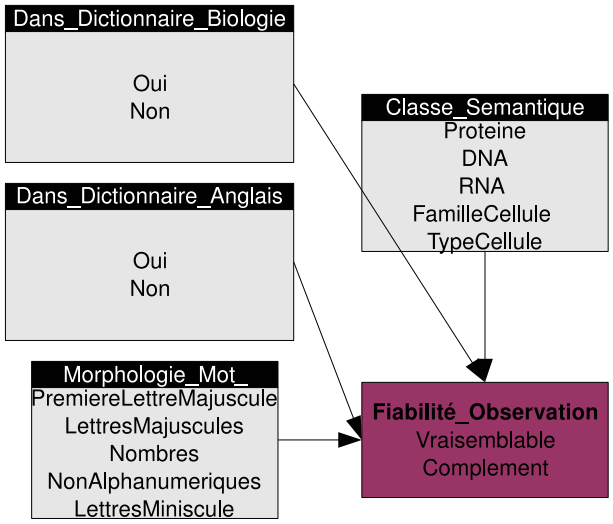

FIG. 7.8 – Attributs renforçant l'annotation des entités nommées de la génomique

nœuds `Fiabilite_Observation` et `SN_Position_Phrase`.

Les probabilités *a priori* du nœud `Fiabilite_Observation` peuvent être apprises grâce aux fréquences d'un corpus d'entraînement mentionnant les rôles grammaticaux des candidats calculés automatiquement et leurs corrections, plus simplement, elles peuvent être fournies par un expert à partir d'une estimation de la qualité de l'annotation réalisée sur un échantillon, ou encore des performances publiées pour l'outil utilisé.

Douter des entités nommées

Une seconde façon de s'assurer de la fiabilité d'une annotation consiste à croiser les informations à notre disposition. Par exemple, pour vérifier qu'un mot ou un GN étiqueté comme EN est correct il suffit souvent de contrôler que le mot ou le GN n'appartient pas au dictionnaire de la langue courante. Une telle vérification aurait considérablement diminué la fiabilité des annotations spécifiant que *not* ou *similar* étaient des noms de protéines.

Pour améliorer la confiance de l'étiquetage des entités nommées nous enrichissons la représentation des entités nommées : elle doit être induite d'une ontologie du domaine du corpus. A titre d'exemple nous reprenons les classes reconnues par l'annotateur d'entités nommées du projet *Genia*[82], voir le nœud `Semantic_Class` de la figure 7.8. Pour ce domaine, des ressources spécialisées, comme des dictionnaires de noms de gènes ou de protéines, existent dont nous pouvons nous servir pour confirmer l'annotation de l'étiqueteur d'entités nommées. Comme le souligne [Proux *et al.*, 1998], les entités nommées de la génomique sont rarement des mots de la langue courante anglaise. La non-appartenance d'un mot de l'entité nommée dans un tel dictionnaire renforcerait son statut. L'étude des caractères non-alphanumériques contenus dans le mot jouerait un rôle identique.

[82] `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/`

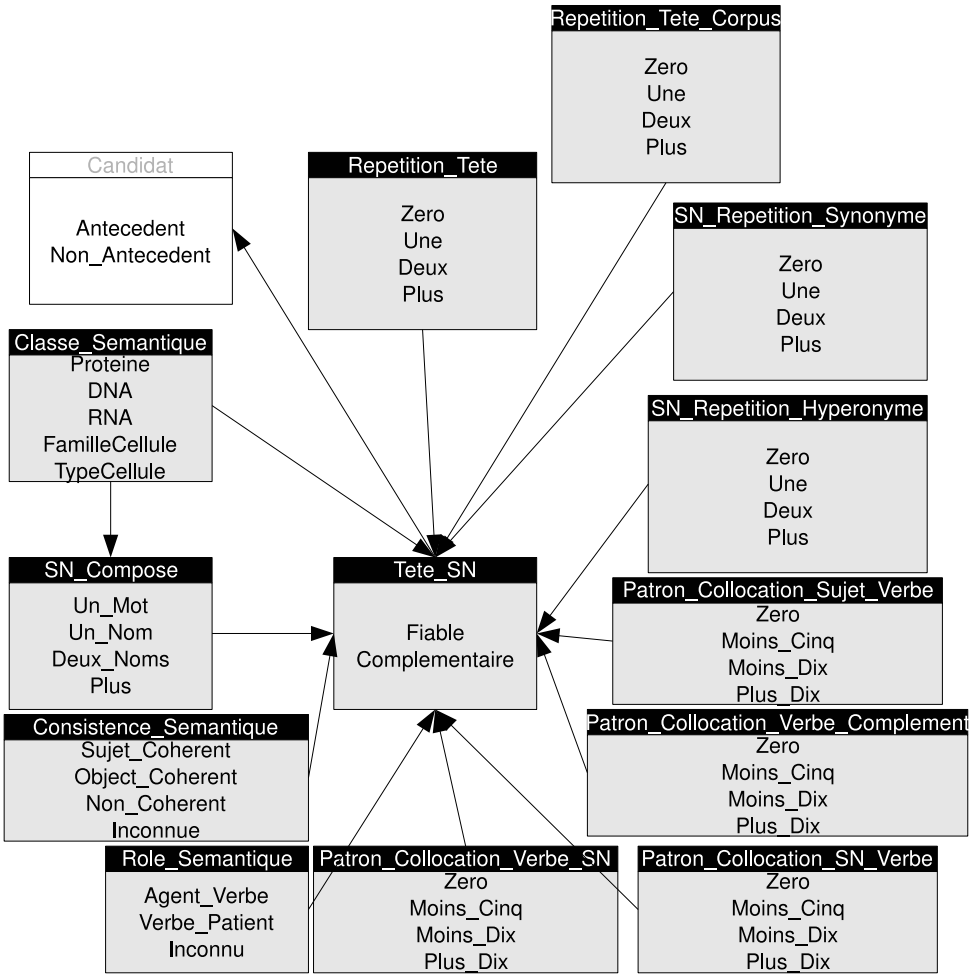

FIG. 7.9 – Attributs renforçant l'annotation de la tête du candidat

Douter de la tête d'un syntagme nominal

Enfin, pour la tâche particulière de la résolution des anaphores, nous avons croisé les informations à notre disposition pour renforcer notre confiance dans la tête du candidat que nous avons calculée.

La tête du candidat reste impliquée dans le calcul d'un nombre important d'attributs. Un attribut très simple, reposant sur le nombre de noms qui composent le candidat, permet de nuancer la fiabilité de la tête calculée, voir la figure 7.9. Si la tête que nous venons de calculer est composée d'un mot qui n'est pas un nom commun ou un nom propre du domaine (par exemple un nom de gène), la tête est très peu fiable et nous souhaitons diminuer la probabilité pour ce candidat d'être choisi comme antécédent. D'autres facteurs peuvent réhausser la fiabilité de la tête du candidat comme la répétition de la tête dans le document[83], dans le corpus[84] ou encore

[83]Ce que nous marquons par une liaison entre le nœud `Tete_SN` et les nœuds `Repetition_Tete`, `SN_Repetition_Synonyme` et `SN_Repetition_Synonyme`. Nous ignorons le nœud `SN_Repetition exacte` car nous recherchons uniquement les répétition de la tête du candidat.

[84]Ce que nous marquons par une liaison entre le nœud `Tete_SN` et les nœuds de collocations et

dans une base de rôles sémantiques[85]. Considérons par exemple les candidats *alpha-amylase synthesis* et *alpha-amylase synthesis in Bacillus licheniformis CCM 2205 were*. Notre système calcule incorrectement la tête du premier candidat *alpha-amylase synthesis* et incorrectement la tête du second *Bacillus licheniformis CCM*. Or la tête *Bacillus licheniformis CCM* n'est répétée qu'une fois dans le même document (dans la phrase suivante) alors que la tête alpha-amylase synthesis est répétée 4 fois dans le corpus (dont une fois dans un document différent). Dans ce cas, les attributs `Repetition_Tete` et `Repetition_Tete_Corpus` diminuent la fiabilité du second candidat et augmentent celle du premier permettant à notre système de privilégier le premier candidat par rapport au second. Pour établir les probabilités *a priori* de la fiabilité de la tête d'un candidat l'interrogation d'un expert semble être la solution la plus simple.

7.3 Améliorer le Classifieur Bayésien pour la classification des pronoms impersonnels

Même si le CB pour la classification des pronoms impersonnels obtient de bonnes performances, nous pouvons encore lui apporter certaines modifications que nous soulignons par la couleur bleue dans la figure 7.10

7.3.1 Douter du rôle grammatical du pronom

La première modification consiste à relativiser la fiabilité du rôle grammatical du pronom. La fiabilité du rôle grammatical du pronom peut être conditionnée par différents attributs, qui sont les mêmes attributs qui influent sur la fiabilité du rôle grammatical du candidat.

7.3.2 Renforcer la fiabilité du pronom anaphorique

La seconde modification est l'ajout des liens de dépendance entre le statut du pronom et les patrons de collocations des candidats à l'antécédence. Pour simplifier la représentation nous ne conservons que les liens vers les patrons d'un unique candidat et nous ignorons les liens de dépendances des patrons de collocations. Le schéma 7.11 de la dernière sous-section montre une vue d'ensemble du réseau. La raison de cet ajout est un renforcement qui n'avait pas été pris en compte dans le précédent réseau. Lorsqu'une collocation identique est trouvée entre le pronom et un candidat, la probabilité du candidat d'être l'antécédent augmente mais aussi celle du pronom d'être anaphorique. La collocation d'un pronom impersonnel ne devrait, en théorie, jamais être partagée par un candidat.

`Repetition_Tete_Corpus`.

[85] Ce que nous marquons par une liaison entre le nœud `Tete_SN` et les nœuds de consistance sémantique et du rôle sémantique.

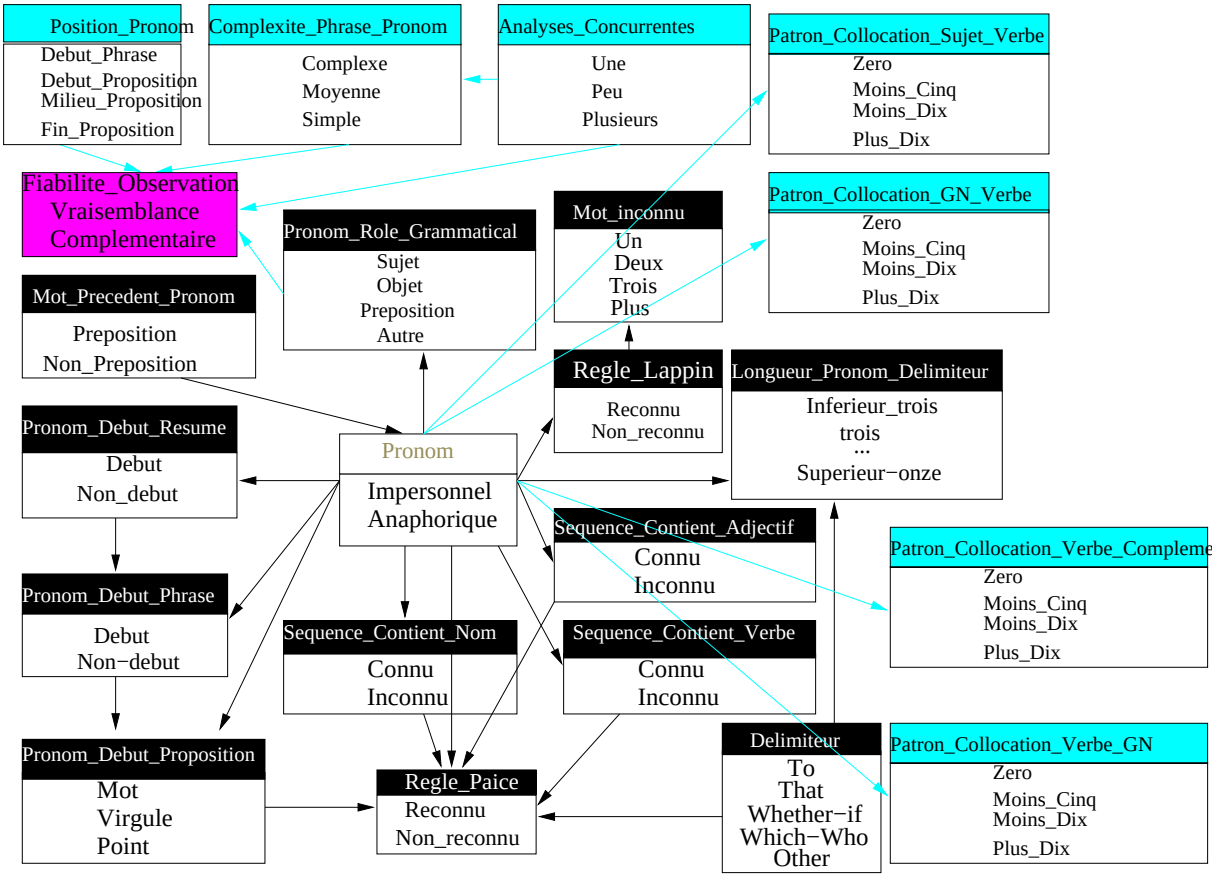

FIG. 7.10 – Correction du CB filtrant les pronoms impersonnels

7.4 Vers un réseau bayésien dynamique pour la résolution des anaphores

La dernière modification majeure du système consiste à passer à un réseau bayésien dynamique (RBD). La structure du *CB* a été volontairement simplifiée pour faciliter l'implémentation et l'analyse des résultats, cette structure modélise imparfaitement notre problème. Les attributs qui composent le réseau bayésien responsable de la classification des pronoms impersonnels ne sont pas indépendants des attributs employés pour la résolution. Les deux réseaux qui travaillaient indépendamment jusqu'à présent doivent être réunis en un unique modèle. De même, le modèle des réseaux bayésiens nous impose de réaliser le classement des candidats de manière indépendante. Lorsqu'un lecteur résout une anaphore, il ne considère pas les candidats isolément mais il tient compte de l'ensemble des candidats pour arrêter son choix. Pour rendre compte de cette interdépendance des candidats nous devons changer de modèle au profit d'un réseau bayésien dynamique. Le nombre de candidats en compétition ne pouvant être connu à l'avance, le réseau doit être créé dynamiquement par le système, chaque *"instant"* du réseau modélisant les caractéristiques d'un candidat.

Nous présentons schématiquement le réseau dans la figure 7.11. Le sous-réseau chargé de la distinction des pronoms impersonnels et anaphoriques a une unique occurrence dans le RBD. Son nœud de prédiction `Pronom` est relié directement au nœud de prédiction `Candidat` des sous-réseaux en charge du calcul de la probabilité d'un candidat d'être l'antécédent. Le nœud `Pronom` remplace le nœud `Impersonal_Filter` du réseau de classification des candidats. Pour chaque candidat de la fenêtre de recherche nous instancions un réseau de résolution et relions les deux nœuds `Candidat` entre le premier candidat lu et son successeur. Nous gardons l'ordre de lecture comme première hypothèse de travail.

La réunion des réseaux augmente la complexité de l'inférence mais elle offre en contrepartie la possibilité de tenir compte de l'ensemble des collocations des candidats dans la classification du pronom et de l'interdépendance des candidats dans le choix de l'antécédent. L'expression de ces nouvelles contraintes devraient améliorer conjointement la classification des pronoms impersonnels et le choix de l'antécédent.

7.5 Conclusion

Dans ce dernier chapitre nous avons proposé une première série de modifications pour améliorer le mécanisme d'inférence :

- diverses corrections de la structure et des attributs du réseau pour augmenter le renforcement entre les attributs ;
- une étape d'apprentissage des paramètres plus complexe où les données corrigées et complétées qui ont été acquises lors de cette expérience seront employées comme données d'expert dans nos futures expériences ;
- une mesure plus précise de la fiabilité des observations des valeurs des attributs en nous appuyant sur les annotations produites par les outils de la plateforme et de nouveaux indices de surface introduits pour estimer la difficulté du calcul automatique des attributs ;
- la modélisation d'un réseau bayésien dynamique réunissant les réseaux bayésiens pour la

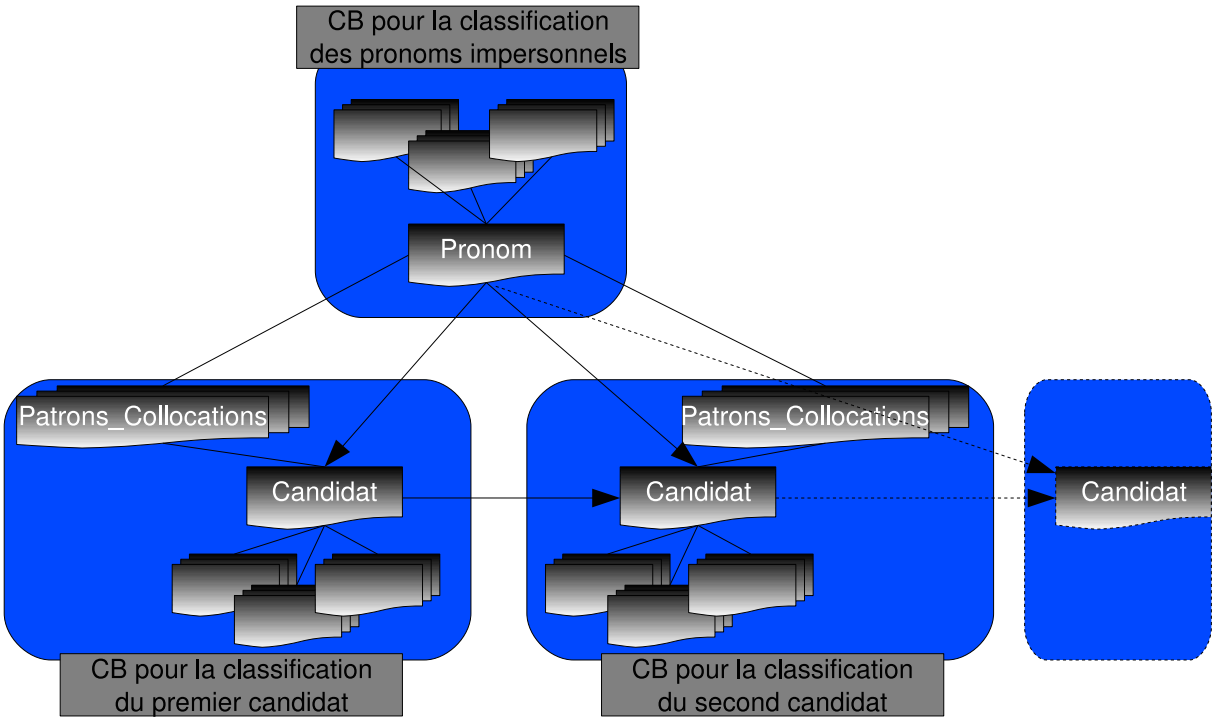

FIG. 7.11 – Un réseau bayésien dynamique pour la distinction des pronoms anaphoriques et leur résolution

classification des pronoms impersonnels et intégrant l'interdépendance entre les candidats lors du choix de l'antécédent ;

La seconde série de modifications vise à enrichir notre système. L'analyse détaillée des erreurs du système *CB* a montré que les connaissances exploitées par notre système étaient encore insuffisantes. Or ces connaissances sont maintenant, en partie, disponibles. Les ressources générales ou spécialisées au domaine, telles que *Wordnet* ou *Gene Ontology* par exemple, fournissent des connaissances sémantiques et ontologiques nécessaires à la résolution des anaphores que nous n'avons pas encore utilisées. La mise à disposition de ces nouvelles connaissances plus complexes rend possible un calcul plus précis de l'élément central et devrait améliorer la stratégie de notre système. La connaissance du centre peut renforcer le poids du candidat le plus saillant lorsqu'ils sont identiques ou l'affaiblir s'ils sont différents.

Toutes ces modifications nous ont servi de base pour concevoir un nouveau système de résolution d'anaphores reposant sur un réseau bayésien dynamique. Nous avons présenté le prototype de ce nouveau système dans la dernière section de ce chapitre. Ce prototype est encore à l'état d'ébauche. Nous devons encore étudier les nouvelles ressources sémantiques et ontologiques que nous souhaitons utiliser pour finaliser la modélisation du nouveau réseau. Nous réaliserons ensuite une série d'expériences identiques à celles que nous avons mené dans le chapitre précédent sur les mêmes corpus pour éprouver nos modifications avant d'évaluer notre nouveau système sur des corpus de tailles plus importantes.

Chapitre 8

Conclusion et perspectives

8.1 Conclusion

L'annotation d'un corpus est une tâche incrémentale que nous réalisons, sans intervention humaine, encore imparfaitement. Les annotations d'entrée qui sont souvent imprécises, erronées ou même manquantes faussent l'inférence du système et détériorent la qualité des annotations qu'il doit ajouter. En exprimant la tâche d'annotation comme un problème de classification nous avons mis en évidence que cette imperfection des annotations est une réalité dont nous devons tenir compte lors de la conception du système.

En raison de la complexité et de la diversité de la langue naturelle, le biais de représentation est élevé, il est difficile d'identifier quelles sont les annotations d'entrée discriminantes qui sont nécessaires pour annoter un texte. De plus ces annotations d'entrée, résultantes elles-même d'un calcul automatique, sont imparfaites *i.e.* imprécises, bruitées ou manquantes. Enfin, l'erreur d'estimation introduite par la construction du corpus sur lequel ont été calculés les poids relatifs des différentes annotations implique une variation importante de leurs fiabilités lorsque le système est appliqué sur un corpus de nature différente.

Nous avons proposé un classifieur reposant sur le modèle des réseaux bayésiens. Ce modèle probabiliste, encore peu employé pour le TAL, est adapté pour classer les données de la langue naturelle.

- La possibilité de représenter dans une formalisation unique des attributs hétérogènes diminue le biais de représentation. Tous les attributs pertinents pour une tâche d'annotation peuvent être exploités lors de l'inférence et les données de chaque classe discriminées au mieux.
- Les probabilités conditionnelles *a priori* associées au réseau expriment un ensemble de contraintes dont nous nous servons pour estimer les valeurs des attributs inconnus et renforcer la fiabilité des valeurs observées prises par les attributs les plus bruités grâce à celles prises par les attributs les plus sûrs.
- L'apprentissage automatique des probabilités conditionnelles est un mécanisme simple pour atténuer l'erreur d'estimation. L'étape d'apprentissage adapte les probabilités conditionnelles au corpus et garantit que les attributs engagés dans la décision sont toujours discriminants.

Nous avons validé notre modèle sur le problème de la résolution du pronom *it* anapho-

rique dans les textes anglais. Pour traiter ce problème on distingue habituellement deux tâches d'annotations complémentaires, la distinction des pronoms impersonnels et anaphoriques et la résolution de la relation anaphorique. La tâche d'annotation des pronoms impersonnels est une tâche plus simple que la tâche de résolution des relations anaphoriques, les attributs discriminants sont moins nombreux, en parti connus et ils se calculent avec une certaine fiabilité. Nous avons conçu et implémenté un classifieur pour chacune des deux tâches et évalué les deux classifieurs sur des corpus de domaines et de genres différents.

L'analyse des résultats du classifieur pour la distinction des pronoms impersonnels a mis clairement en évidence l'intérêt de notre approche. Les résultats de notre classifieur sont meilleurs que ceux des systèmes concurrents de l'état de l'art.

Sur la tâche de la résolution de la relation anaphorique, les résultats du second classifieur sont moins satisfaisants mais ils restent comparables aux résultats des systèmes de l'état de l'art les plus connus. L'analyse de nos résultats a permis de mettre en avant trois points intéressants :

- L'intégration au sein du réseau d'un classifieur des connaissances linguistiques complexes dont les valeurs des attributs sont bruitées ou manquantes, ne dégrade pas fortement ses performances, en comparaison de celles d'un classifieur qui exploiterait uniquement les indices de surfaces plus fiables.
- Il existe un seuil de fiabilité pour les connaissances linguistiques complexes au dessus duquel leur apport est significatif (même si ces connaissances soit encore incertaines).
- Les probabilités que le système associe à chaque candidat les ordonnent partiellement. Or, l'antécédent, lorsqu'il n'est pas en première position, apparaît souvent en deuxième position. Ce fait nous semble important pour la conception d'un système de résolution semi-automatique performant.

Les performances du classifieur que nous avons conçu pour la résolution des relations anaphoriques montrent qu'il est encore trop simple. Nous avons proposé différentes corrections de la structure du réseau et de son paramétrage. Ces corrections devraient améliorer le mécanisme d'inférence du classifieur qui s'est révélé peu robuste au bruit et au silence des attributs. Nous avons aussi suggéré d'enrichir notre classifieur de nouvelles connaissances sémantiques et ontologiques indispensables dans la recherche de l'antécédent. L'intégration de ces modifications à notre modèle nous a conduit à la conception d'un nouveau classifieur reposant sur un réseau bayésien dynamique dont nous avons esquissé le prototype dans la dernière section de cette thèse.

8.2 Perspectives

Notre travail ouvre deux pistes de réflexions différentes. La première est spécifique à la résolution des anaphores, la seconde, plus générale, concerne l'étude de la prise en compte de l'imperfection des annotations utilisées par les systèmes de TAL.

Nous avons validé notre approche sur la tâche de distinction des pronoms impersonnels mais pas sur la tâche de la résolution des relations anaphoriques, nos résultats ne sont pas significativement meilleurs que ceux des systèmes de l'état de l'art. En corrigeant notre classifieur nous avons conçu un classifieur plus complexe. Nous souhaitons achever prochainement le prototype de ce classifieur puis l'implémenter. Afin de confirmer une amélioration de nos résultats sur la tâche de résolution, nous testerons ce classifieur avec un protocole identique à celui utilisé pour

notre premier classifieur.

Les structures des réseaux bayésiens que nous avons étudiées, ainsi que celle du prototype, ont toujours été imposées par un expert. Les travaux de [Bouckaert, 2002] ont montré que l'apprentissage de la structure est non seulement possible mais aussi améliore le réseau. Après avoir corrigé nos données, nous pourrons employer des algorithmes d'apprentissage de la structure de nos classifieurs et soumettre les réseaux obtenus aux linguistes. Nous espérons valider les liens d'influence de nos réseaux et découvrir de nouveaux liens de renforcement pertinents.

Nous avons concentré notre travail sur la modélisation du raisonnement à partir d'annotations imparfaites et mis temporairement de coté l'exigence de résultat. Dans la section 4.4 nous avons succintement présenté les modèles discriminants. Ces modèles, et plus particulièrement les séparateurs à vaste marges, sont généralement plus performants que les modèles génératifs. Nous envisageons de transformer nos classifeurs en modèles discriminants et espérons une augmentation des résultats sur les tâches d'annotation des pronoms impersonnels et de résolution des anaphores.

En l'état actuel, les performances des systèmes automatiques de résolution des anaphores sont encore trop faibles pour être une aide significative à un système de TAL. Nous devons donc mener en parallèle des recherches pour la résolution automatique des anaphores et une réflexion sur l'intégration d'un module de résolution semi-automatique des anaphores au sein d'un système de TAL. Une évaluation de l'amélioration des performances du système de TAL relative à la pénibilité d'utilisation du module de résolution confirmera la pérenité du projet.

Même si un système de résolution automatique n'est pas une aide significative, selon [Mitkov *et al.*, 2007], il améliore quelque peu les performances d'un système de TAL. Nous allons intégrer notre système dans la plateforme Ogmios. Nous pourrons confirmer les précédents résultats sur une tâche de recherche d'information réalisée sur de gros corpus scientifiques.

D'un point de vue plus général, nous poursuivons notre travail sur la modélisation et l'apport des annotations imparfaites.

Nous avons établi l'existence d'un seuil de qualité des annotations en dessous duquel les annotations sont inutiles lors de l'inférence. Des expériences supplémentaires doivent encore être menées pour mesurer ce seuil. Nous possèderons alors un outil pour apprécier l'utilité d'une information pour une tâche de TAL : si la qualité de l'annotation automatique de cette information est inférieure au seuil mesuré, il est inutile de l'ajouter aux données d'entrée du système, elle n'améliorera pas ses performances. Inversement, si sa qualité est supérieure au seuil, nous savons qu'elle sera profitable au système et nous pouvons chercher à en mesurer son impact.

Dans un travail futur, nous avons pour objectif d'intégrer, en plus de l'incertitude et de l'absence d'une annotation, son imprécision dans le processus de raisonnement d'un système de TAL. Cette intégration peut se révèler profitable comme le montrent [Witte & Bergler, 2003] avec l'évaluation d'un algorithme de résolution des chaînes de coréférences qui regroupe les coréférents dans des ensembles floues.

Enfin, les travaux de l'état de l'art que nous avons cités dans la section 4.5 et nos résultats sur la distinction des pronoms impersonnels ont montré l'intérêt de la modélisation de l'imperfection des annotations pour des tâches d'annotations simples. Nous avons choisi la tâche de résolution des pronoms anaphoriques car cette tâche complexe se prêtait à notre étude. Mais le choix d'autres tâches complexes, telles que l'extraction de relations spécifiques, l'analyse syntaxique ou encore le résumé automatique, pourraient être considérés et consolider nos résultats.

Annexes

1 Annexe 1

Analyse détaillée des FP pour la $1^{ère}$ itération
Délimiteur to
Cause : 2 erreurs d'annotation Phrases : - Also, the new vector makes it feasible to generate P1 libraries from small amounts of genomic insert DNA, such as from sorted chromosomes. - Recent advances in DNA technology have made it possible to analyse the structure and function of the replication origin region of the chromosomes of various bacteria. Cause : 2 Automates imprécis Phrases : - It has been shown to possess anti-inflammatory activity in a variety of animal models and more recently to inhibit IL-2 induced signal transduction. - Although Vpr is not required for viral replication in transformed cell lines and primary T lymphocytes, it is essential for productive infection of macrophages and monocytes and appears to be important for pathogenesis in vivo. Cause : 6 Descriptions non discriminantes Phrases : - It did not replicate independently in Bacillus subtilis, but *it* could be made to integrate into the chromosome of B. subtilis if sequences homologous to chromosomal sequences were inserted into it. - However, it was very homologous to the promoter sequence of the spo0B gene from B. subtilis. - It is highly homologous to alpha(1)-antitrypsin (AT). - Then, it was used as a bait to screen the human fetal liver cDNA library by yeast two-hybrid, with the cDNA fragment inserted into pACT2 vector and fused in-frame to the Gal4 activation domain. - It becomes restricted to the central nervous system during the second half of gestation. - The XPD gene was initially named ERCC2 (excision repair cross complementing) as it was cloned using human DNA to complement the ultraviolet sensitivity of a rodent cell line.

Analyse détaillée des FP pour la $1^{ère}$ itération
Délimiteur that
Cause : 1 automate imprécis Phrases : - The results of electrophoretic mobility shift assays using DNA with 5'-(CGG)(n)-3' repeats of different lengths render it likely that oligomers of the p20 protein bind to the repeat.
Délimiteur whether_if
Cause : 1 Description non discriminante Phrases : - It also discusses new insights into how the structure of these tissues supports effective immune responses. Cause : 1 Configuration spécifique Phrases : - It did not interfere with the activity of the enzymes for sulfate assimilation but clearly inhibited the induction of those enzymes when S. cerevisiae cells were transferred from rich medium to a synthetic methionine-free medium. Cause : 1 Transducteur imprécis Phrases : - Although in silencing competent cells there was no correlation between the fraction of cells in which a telomeric gene was repressed and the fraction of cells in which it was localized to the periphery, no condition was found where the telomere was both silenced and away from the periphery.
Délimiteur who_which
Cause : 1 Transducteur imprécis Phrases : - A stuttering or slippage model to account for these events has been further refined to include a pressure which displaces the nascent strand in a given direction once it has disengaged from the template, and the similarities of this model to those which account for readthrough of cellular RNA polymerase transcription blocks are discussed. Cause : 4 Configurations spécifiques Phrases : - Importantly, it was also present in the nucleus of HeLa cells, which lack HNF1, and in the cytoplasm of fibroblasts that have little or no tetrahydrobiopterin. - However, it should be distinguished from poorly differentiated thyroid carcinoma, which has a reported lower survival rate compared with the solid variant of papillary carcinoma. - In the present study, the AML1/ETO mRNA could be detected by RT-PCR in bone marrow (BM) and/or peripheral blood (PB) samples from all 18 patients who had been maintaining complete remission for 12 to 150 months (median, 45 months) following chemotherapy or PB stem cell transplantation (PBSCT), whereas it could not be detected in four patients who had been maintaining remission for more than 30 months following allogeneic BM transplantation (BMT). - The retinoblastoma protein (Rb), a key regulator of cell cycle progression, can bind the transcription factor E2F converting it from a positive transcriptional factor capable of driving cells into S phase into a negative complex which arrests cells in G1.

TAB. 1 – Détail des FP produit par le CB sur le corpus Génétique

Analyse détaillée des FN pour la $1^{ère}$ itération

Aucun délimiteur

Cause : 2 Délimiteurs implicites

Phrases :

- The NUP98-HOXA9 fusion transcript was detected by RT-PCR, suggesting its role in the malignant transformation as it has been postulated for other t (7 ;11) -associated leukemias.
- The expression of 3 alternatively spliced isoforms of Ang-1 was confirmed by RT-PCR using specific primer pairs derived from junction sites and the 3' end of Ang-1 cDNA, and it was further demonstrated by nuclease protection assay, Northern blotting, and immunoblotting in CHRF cells.

Cause : 4 Séquences reconnues uniquement par les automates

Phrases :

- It remains to be shown which molecular events lead to the specific 'pre-activation', i.e. constitutive nuclear translocation and DNA binding, of these members of NF-AT, NF-kB and AP-1 factor families.
- Since medical treatments or stress commonly increase opioid levels, it is important to understand the mechanisms by which opioids affect T lymphocyte functions.
- However, it remains to be determined whether the differential transcriptional activity of MCAT elements in SM vs. non-SM is due to differences in expression of TEF-1 or TEF-1-related proteins or to unique (cell type specific) combinatorial interactions of the MCAT elements with other cis-elements and trans-factors.
- However, using fluorescence in situ hybridization (FISH) and by screening for the TEL/AML1 rearrangement by the polymerase chain reaction (PCR), it has been demonstrated to be the most frequent known structural chromosomal abnormality in childhood acute lymphoblastic leukemia (ALL).

Cause : 1 Séquence ignorée par les règles et les automates

Phrases :

- It is necessary, therefore, to understand the degree to which xenobiotics can disrupt endocrine systems.

Cause : 7 Erreurs d'annotation

Phrases :

- However, it has marked differences in that the promoter contains Egr-1 sites and lacks both hypoxia-inducible factor-1 and AP-1 sites.
- It is predicted to be a transcription factor as it contains six Kruppel-like Zinc finger motifs and a N-terminal BTB/POZ domain, a protein/protein interaction interface that is widely conserved in Metazoans.
- Determining the sub-nuclear localization of proteins is therefore important for understanding genome regulation and function, and it also provides clues to function for novel proteins.
- It has been shown to enhance the nuclear transport of the HIV-1 pre-integration complex, activate transcription of cellular and viral promoters, and arrest the cell cycle at the G2/M check-point.
- It has been shown to both repress transcriptional activators by direct protein-protein interactions and to bind DNA hairpin structures and repress target genes.

Analyse détaillée des FN pour la $1^{ère}$ itération
Aucun délimiteur
Phrases : - It has been shown to be associated with a B-precursor phenotype and an excellent prognosis. - IEP72 often synergistically affects the activation by IEP86, although it has not previously been shown to directly interact in vitro with IEP86, TBP, or transcription factors (e.g., Sp1 and Tef-1) bound by IEP86.
Délimiteur to
Cause : 8 Descriptions non discriminantes Phrases : - We further demonstrate that by using cDNA hybrid selection, it is relatively straightforward to isolate cDNAs that correspond to genes embedded in the EVI-1-binding sublibrary. - Looking back at successes and failures in newer approaches to treating IBD, it is tempting – although still difficult – to draw conclusions about pathogenesis. - It was identical to GAS41, a sequence amplified in human gliomas. - Gel-shift analysis shows that protein binds specifically to this region, but competition studies suggest that it is unlikely to be LBP-1. - In the peripheral tissues that express CRH mRNA it will be very interesting to document the specific cell type of synthesis by using combined immunocytochemical and in situ histochemical techniques. - However, although constitutive activation of NF-kappaB efficiently induced promoter activity, it was not sufficient to induce either ICAM-1 mRNA or ICAM-1 protein. - Given the induction of expression and regulation of the GATA-4 and GATA-6 zinc finger family of transcription factors in the gonads by gonadotropins, it was in our interest to explore their expression in the adrenals. - Since chromosome number abnormalities may be associated with submicroscopic gene rearrangements, it should be important to search for them for a better understanding of mechanisms of leukemogenesis, and to understand the prognostic heterogeneity in leukemic patients with aneusomies without apparent chromosome structure rearrangements. Cause : 1 Transducteur imprécis Phrases : - These studies suggest that it is possible to create chimeric transcription factors able to strongly and selectively activate genes downstream of p53. Cause : 2 Configurations spécifiques Phrases : - It has long been the goal of molecular biologists to design DNA-binding proteins for the specific control of gene expression. - It is therefore of great interest for clinical laboratories to have alternative technical possibilities for the set-up of standardized molecular tests.

Analyse détaillée des FN pour la $1^{ère}$ itération

Délimiteur that

Cause : 2 Appositions

Phrases :

- It was further observed , by Western blot analysis, that the levels of Sp1 protein are higher in PC-3 cells when compared to levels in HS27 cells, possibly contributing to a tissue-specific effect.
- It has been demonstrated, therefore, that the cAMP-dependent regulation of the P-450 (SCC) gene in adrenal cortex is faithfully reflected in the transient expression system using Y-1 cells and the fusion gene and that a cis-acting DNA element (s) in response to cAMP is present within the 5'-flanking sequence (5.4 kb) of the P-450 (SCC) gene.

Cause : 6 Descriptions non discriminantes

Phrases :

- It was striking that there was no apparent consensus gene end sequence between the M and the F genes and that the M gene was transcribed exclusively as a dicistron with the F gene.
- It is envisioned that transformation by the EWS/ATF1 fusion protein results from aberrant transcriptional regulation of genes that are normally regulated by ATF1.
- It is interesting to note that acute ethanol re-challenge immediately after chronic treatment did not result in RACK1 translocation to the nucleus, and nuclear compartmentalization of RACK1 in response to acute ethanol was detected only after 24 hr of withdrawal.
- It is the hope that what we have learnt from APL will benefit further developments of anti-leukemia therapy.
- It is remarkable that in B-CLL cells the nuclear appearance and DNA binding of specific transcription factors is dramatically affected whereas other members of the same factor family remained unaltered in these leukemic cells.
- It is possible that the RT activity was involved in the spread of this major class of retroelements by retrotransposition, and in fact it cannot be excluded that this retrovirus group is still mobile.

Cause : 1 Configuration spécifique

Phrases :

- It was shown in an in vivo experiment that treatment of rats with thyroid hormone increased hypothalamic OT mRNA levels, the pituitary OT content, as well as OT levels in blood.

Cause : 2 Ressources incomplètes

Phrases :

- If this gene should turn out to be involved in the pathogenesis of lung and kidney tumors, it will indicate that transmembrane protein tyrosine phosphatase may represent a class of tumor suppressors.
- Although Pol II similarly interacts with TFIIB, it is notable that C17 has no similarity to any Pol II subunit.

Cause : 1 Transducteur imprécis

Phrases :

- It has not been possible to isolate mutations mutations.

Analyse détaillée des FN pour la $1^{ère}$ itération
Délimiteur whether_if
Cause : 2 Descriptions non discriminantes Phrases : - It is not well understood, however, whether Shh signalling also controls the activities of Gli proteins post-translationally and whether these activities have activating or repressing effects on target genes in vivo. - It was our aim to find out whether MLL mutant (MLLmu) and MLL wild-type (MLLwt) acute leukemia-derived cell lines might likewise be discriminated on the basis of HOX gene expression.

TAB. 2 – Détail des FN produit par le CB sur le corpus Génétique

2 Annexe 2

Analyse détaillée des FP
Cause : 7 Mauvais paramétrage Phrases : - [It] provides such features as a box around the pointer which makes it easier to locate. - In fact, once Emacspeak is installed and running, it provides a fluent interface to the rich set of online documentation including the info pages, and makes learning what you need a lot easier. - The main thing about patch is that program understands this, and that it knows how to guess what to do when the Linux developers change things in that file. - Although, [it] only acts as a plain text reader (similar to IBM's one for the PC) when controling programs it doesn't understand, with those that it does, it can provide much more sophisticated control. - Once this VCR is connected to an aerial, it will be tuned in and the correct date and time will be set automatically as soon as the mains lead is connected to the mains. - Once the VCR is connected to an aerial, it will be turned in and the correct date and time will be set automatically as soon as the mains lead is connected to the mains. - Whether you're at home or on the road, you can connect to your office computer and the network it is on by using a modem and dialing in to a remote-access server on your network.
Cause : 1 Erreur d'automate Phrases : - However, there seem to be interactions between xpuff and the window manager which could make it difficult to use.
Cause : 2 Descriptions non discriminantes Phrases : - It may be used to produce large print documents of almost any nature. - The EULA is the contract regarding your use of the licensed product and it grants you a specific right to use the Microsoft software on your computer.

TAB. 3 – Détail des Faux Positifs sur le corpus MARS

Analyse détaillée des FN
Cause : 5 Absences de ressource Phrases : - There are also others which it is worth researching which cover computer use more generally. - It is recommended that you also create a symbolic link to the CD-ROM device to make [it] easier to remember. - It is intended that this advise is used in conjunction with the instruction books for your ancillary equipment. - The look and feel of Windows has been improved to make it easier and faster for you to get your work done. - When you do view the list of messages, you can expand and collapse individual topics to make it easier to look for messages or topics of interest.
Cause : 1 Forme interrogative non-gérée Phrases : - What does it mean when I get a kernel message from the IDE CD-ROM driver like ""hdxx :
Cause : 2 Délimiteurs implicites/absents Phrases : - Grade II Braille is difficult to learn, but is almost certainly worth [it] since it is much faster. - It is worth having more than one size or a good quality set with interchangeable bits.
Cause : 2 Délimiteurs non-gérés Phrases : - Where possible, make it clear what different parts of your program are what. - It is not suitable for sheet metal, unless you are making small straight cuts at the edges.
Cause : 1 Transducteur imprécis Phrases : - [It] will take a while to search for the page, once it is found it will be displayed on the screen.

TAB. 4 – Détail des Faux Négatifs sur le corpus MARS

Détail des erreurs humaines corrigeables
Cause : 1 Annotation contestée Phrases : - If your software is only usable through a graphical interface then [it] can be very hard to make it usable for someone who can't see.
Cause : 3 Anaphores causales annotées impersonnelles Phrases : - In principle it should be possible to put together a complete, usable Linux system for a visually impaired person for about $500 (cheap and nasty PC + sound card). I doubt [it] would work in practice because the software speech synthesisers available for Linux aren't yet sufficiently good. - No doubt you made a contribution and I haven't mentioned it. Don't worry, [it] was an accident. - Always use tools correctly. If [it] feels very awkward, stop.
Cause : 4 Anaphores verbales annotées impersonnelles Phrases : - If someone gets this to work properly, please send me the details of how you did it. - This could be changed if anyone with a basic level of kernel programming ability wants to do it. - Even if the author of the software is willing to make exceptions, it makes the user vulnerable both to changes of commercial conditions (some company buys up the rights) and to refusal from people they could work for (many companies are overly paranoid about licenses). - It is not possible to give safety guidelines to cover every do [it] yourself manual.
Cause : 4 Mauvais traitements d'un caractère non-alphanumerique Phrases : - I have returned this patch to the maintainer of NFBtrans and he says that he has included it, so if you get a version later than 740, you probably won't have to do anything special. - If you can read from the drive but cannot mount it, first verify that you compiled in ISO-9660 file system support by reading/proc/filesystems, as described previously. - Most of the time, it's possible to replace this by making the entire screen (or terminal emulator) flash. - If it's easy to change fonts then people will be able to change to one they can read.
Cause : 3 Erreurs d'implémentation divers Phrases : - This TV set is capable of receiving German TOP Teletext transmissions, it will automatically switch to them if a German broadcast is being received, for example, from Satellite channels. - Because CD-ROM drives are changing very quickly, it is difficult to list which models support reading digital data. - [It] is recommended that you also create a symbolic link to the CD-ROM device to make it easier to remember.

TAB. 5 – Détail des erreurs humaines corrigeables

Bibliographie

[Alphonse *et al.*, 2004] Erick Alphonse, Sophie Aubin, Philippe Bessières, Thierry Hamon, Gilles Bisson, Sandrine Lagarrigue, Adeline Nazarenko, Alain-Pierre Manine, Claire Nédellec, Mohamed Ould Abdel Vetah, Thierry Poibeau, et Davy Weissenbacher. Event-based information extraction for the biomedical domain : the caderige project. In *Proceedings on International Joint Workshop on Natural Language Processing in Biomedecine and its Applications (BioNLP/LNPBA), International Conference on Computational Linguistics (COLING'04)*, pages 43–49. in N. Collier, P. Rush, A. Nazarenko (eds), 2004.

[Amsili & Bras, 1998] Pascal Amsili et Myriam. Bras. Drt et compositionnalité. *Traitement Automatique des Langues*, 39(1) :131–160, 1998.

[Amsili *et al.*, 2005] Pascal Amsili, Pascal Denis, et Laurent Roussarie. Les anaphores abstraites en français : représentation formelle. *Traitement Automatique des Langues*, 46(1) :15–39, 2005.

[Aone & Bennett, 1995] Chinatsu Aone et Scott Bennett. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Meeting of the Association for Computational Linguistics*, pages 122–129, 1995.

[Aubin & Hamon, 2006] Sophie Aubin et Thierry Hamon. Improving term extraction with terminological resources. In *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, pages 380–387. Springer, 2006.

[Aubin *et al.*, 2005] Sophie Aubin, Adeline Nazarenko, et Claire Nedellec. Adapting a general parser to a sublanguage. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'05)*, pages 89–93, 2005.

[Berroyer, 2004] Jean-François Berroyer. Tagen, un analyseur d'entités nommées : conception, développement et évaluation. Master's thesis, Mémoire de d.e.a d'intelligence artificielle, Université Paris-Nord, 2004.

[Bouchon-Meunier, 1995] Bernadette Bouchon-Meunier. *La logique floue et ses applications*. 1995.

[Bouckaert, 2002] Remco Bouckaert. Low level information extraction, a bayesian network based approach. In *Workshop on Text Learning (TextML-2002)*, 2002.

[Boudreau & Kittredge, 2005] Sylvie Boudreau et Richard Kittredge. Résolution des anaphores et détermination des chaînes de coréférences. *Traitement Automatique des Langues (TAL)*, 46 :41–69, 2005.

[Bourigault & Jacquemin, 2000] Didier Bourigault et Christian Jacquemin. Constitution de ressources terminologiques. In *Ingénierie des langues*, chapter 9, pages 215–233. Hermes Science, 2000. Sous la direction de Jean-Marie Pierrel.

[Boyd *et al.*, 2005] Adriane Boyd, Whitney Gegg-Harrison, et Donna Byron. Identifying non-referential it : A machine learning approach incorporating linguistically motivated patterns. *Traitement Automatique des Langues (TAL), Special issue on Anaphora Resolution*, 46(1) :71–90, 2005.

[Brennan *et al.*, 1987] Susan Brennan, Marilyn Friedman, et Carl Pollard. A centering approach to pronouns. In *Proc. 25th Annual Meeting of the ACL*, pages 155–162, 1987.

[Brewka & Eiter, 2000] Gerhard Brewka et Thomas Eiter. Prioritizing default logic. In *Intellectics and Computational Logic*, pages 27–45, 2000.

[Byron & Gegg-Harrison, 2004] Donna Byron et Whitney Gegg-Harrison. Evaluating optimality theory for pronoun resolution algorithm specification. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2004)*, pages 27–32, 2004.

[Carbonell & Brown, 1988] Jaime Carbonell et Ralf Brown. Anaphora resolution : a multi-strategy approach. *COLING-88*, pages 96–101, 1988.

[Cardie & Wagstaff, 1999] Claire Cardie et Kiri Wagstaff. Noun phrase coreference as clustering. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 82–89, 1999.

[Cohen, 2007] Ariel Cohen. Anaphora resolution as equality by default. In Antonio Branco, editor, *Anaphora : Analysis, Algorithms and applications (DAARC'07)*, pages 45–58, 2007.

[Cornuéjols & Miclet, 2002] Antoine Cornuéjols et Laurent Miclet. *Apprentissage Artificiel*. Eyrolles, 2002.

[Dagan & Itai, 1990] Ido Dagan et Alon Itai. Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, volume 3, pages 1–3, 1990.

[Denoyer & Gallinari, 2004] Ludovic Denoyer et Patrick Gallinari. Bayesian network model for semi-structured document classification. *Information Processing and Management*, 2004.

[Denoyer, 2004] Ludovic Denoyer. *Apprentissage et Inférence Statistique dans les bases de documents structurés*. Phd thesis, University of Paris VI, LIP6, 8 rue du capitaine Scott, 75015 PARIS, December 2004.

[Derivière *et al.*, 2006] Julien Derivière, Thierry Hamon, et Adeline Nazarenko. A scalable and distributed nlp architecture for web document annotation. In *Advances in Natural Language Processing (5th International Conference on NLP, FinTAL 2006)*, pages 56–67, 2006.

[Ding & Peng, 2004] Zhongli Ding et Yun Peng. A probabilistic extension to ontology language owl. In *Proceedings of the 37th Hawaii International Conference on System Sciences*, 2004.

[Domingos & Pazzani, 1996] Pedro Domingos et Michael Pazzani. Beyond independence : conditions for the optimality of the simple bayesian classifier. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 105–112, 1996.

[Evans, 2001] Richard Evans. Applying machine learning toward an automatic classification of it. *Literary and linguistic computing*, 16 :45–57, 2001.

[Fabiani, 1996] Patrick Fabiani. *Représentation dynamique de l'incertain et stratégie de prise d'information pour un système autonome en environnement évolutif*. PhD thesis, Ecole Nationale Supérieure de l'Aéronautique et de l'Espace, 1996.

[Gandon, 2002] Fabien Gandon. Ontology engineering : a survey and a return on experience. Technical report, INRIA, projet Acacia, 2002.

[Ge *et al.*, 1998] Niyu Ge, John Hale, et Eugene Charniak. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 161–170, 1998.

[Grishman, 1997] Ralph Grishman. Tipster architecture design document version 3.1. Technical report, DARPA, 1997.

[Grosz *et al.*, 1983] Barbara Grosz, Aravind Joshi, et Scott Weinstein. Providing a unified account of definite noun phrases in discourse. pages 44–50, 1983.

[Grosz *et al.*, 1995] Barbara Grosz, Scott Weinstein, et Aravind. Joshi. Centering : a framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2) :203–225, 1995.

[Habert *et al.*, 1997] Benoît Habert, Adeline Nazarenko, et André Salem. *Les linguistiques de corpus*. U Linguistique. Armand Colin/Masson, Paris, 1997.

[Halpern, 2003] Joseph Halpern. *Reasoning about uncertainty*. MIT press, 2003.

[Hamon & Nazarenko, 2008] Thierry Hamon et Adeline Nazarenko. Le développement d'une plate-forme pour l'annotation spécialisée de documents web : retour d'expérience. *Traitement Automatique des Langues (TAL)*, 2008.

[Hansson, 2003] Sven Ove Hansson. Ten philosophical problems in belief revision. *Journal of Logic and Computation*, 13(1) :37–49, 2003.

[Hirschman & Chinchor, 1997] Lynette Hirschman et Nancy Chinchor. Muc-7 coreference task definition. version 3.0. Technical report, 1997.

[Hobbs, 1976] Jerry Hobbs. Pronoun resolution. Technical Report Research Report 76-1, New York : Department of Computer Science, 1976.

[Huang, 2000] Yan Huang. *Anaphora : A cross-linguistic study*. Oxford Studies in Typology and Linguistic Theory, 2000.

[Jensen & Nielsen, 2007] Finn Jensen et Thomas Nielsen. *Bayesian Networks and Decision Graphs (Information Science and statistics)*. Springer, 2007.

[Kamp & Reyle, 1993] Hans Kamp et Uwe Reyle. *From discourse to logic*. Dordrecht, 1993.

[Kayser & Levy, 2004] Daniel Kayser et François Levy. Modélisation symbolique du raisonnement causal. *Intellectica*, 38 :291–323, 2004.

[Kayser, 1997] Daniel Kayser. *La représentation des connaissances*. Collection informatique edition, 1997.

[Kehler *et al.*, 2004] Andrew Kehler, Douglas Appelt, Lara Taylor, et Aleksandr Simma. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technology Conference*, pages 289–296, 2004.

[Keller & Lapata, 2003] Frank Keller et Mirella Lapata. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3) :459–484, 2003.

[Lappin & Leass, 1994] Shalom Lappin et Herbert Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4) :535–561, 1994.

[Litran *et al.*, 2004] José Carlos Clemente Litran, Kenji Satou, et Kentaro Torisawa. Improving the identification of non-anaphoric it using support vector machines. In *Actes d'International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 58–61, 2004.

[Manning & Shutze, 1999] Christopher Manning et Hinrich Shutze. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[McCord, 1993] Michael McCord. Heuristics for broad-coverage natural language parsing. In *Proceedings ARPA Human Language Technology Workshop*, 1993.

[Mitkov *et al.*, 2001] Ruslan Mitkov, Branimir Boguraev, et Shalom Lappin. Introduction to the special issue on computational anaphora resolution. *Computational Linguistics*, 27(4) :473–477, 2001.

[Mitkov *et al.*, 2002] Ruslan Mitkov, Richard Evans, et Constantin Orasan. A new, fully automatic version of mitkov's knowledge-poor pronoun resolution method. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, 2002.

[Mitkov *et al.*, 2007] Ruslan Mitkov, Richard Evans, Constantin Orasan, Le An Ha, et Viktor Pekar. Anaphora resolution : To what extent does it help nlp applications ? *Anaphora : Analysis, Algorithms and Applications, Springer*, pages 179–190, 2007.

[Mitkov, 1994] Ruslan Mitkov. An integrated model for anaphora resolution. In *Proceedings of the 15th International Conference on Computational Linguistics COLING'94*, pages 1170–1176, 1994.

[Mitkov, 1997] Ruslan Mitkov. Factors in anaphora resolution : They are not the only things that matter. In *ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 14–21, 1997.

[Mitkov, 1998] Ruslan Mitkov. Robust pronoun resolution with limited knowledge. In *COLING-ACL*, pages 869–875, 1998.

[Mitkov, 2002] Ruslan Mitkov. *Anaphora Resolution*. Longman(Pearson Education), 2002.

[Nazarenko *et al.*, 2006] Adeline Nazarenko, Eric Alphonse, Julien Derivière, Thierry Hamon, Guillaume Vauvert, et Davy Weissenbacher. The alvis format for linguistically annotated documents. In *Proceedings of the Fifth International Conference on Language Resources and Evaluatio (LREC)*, pages 1782–1786, 2006.

[Naïm *et al.*, 2004] Patrick Naïm, Pierre-Henri Wuillemin, Philippe Leray, Olivier Pourret, et Anna Becker. *Réseaux bayésiens*. Eyrolles, Paris, 2004.

[Ng & Cardie, 2002] Vincent Ng et Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, 2002.

[Nouioua, 2007] Farid Nouioua. *Extraction et Utilisation des Normes pour un Raisonnement Causal dans un Corpus Textuel*. PhD thesis, Laboratoire d'informatique de Paris-Nord, Université Paris-Nord, 2007.

[Nugues *et al.*, 2004] Pierre Nugues, Sylvain Dupuy, et Arjan Egges. Carsim : un système pour convertir des textes en scènes tridimensionnelles animées. In *14e congrès de l'AFRIF-AFIA, Reconnaissance des formes et Intelligence artificielle*, volume 2, pages 695–703, 2004.

[Paass, 1988] Gerhard Paass. Probabilistic logic. *Non-Standard Logics for Automated Reasoning*, pages 213–251, 1988.

[Paice & Husk, 1987] Chris Paice et Gareth Husk. Towards the automatic recognition of anaphoric features in english text : the impersonal pronoun it. *Computer Speech and Language*, 2 :109–132, 1987.

[Peshkin & Pfeffer, 2003] Leonid Peshkin et Avi Pfeffer. Bayesian information extraction network. *Proceedings of 18th International Joint Conference of Artificial Intelligence*, 2003.

[Philippe *et al.*, 2003] Besnard Philippe, Fanselow Gisbert, et Schaub Torsten. Optimality theory as a family of cumulative logics. *Journal of Logic, Language, and Information*, 12 :153–182, 2003.

[Ponzetto & Strube, 2006] Simone Paolo Ponzetto et Michael Strube. Semantic role labeling for coreference resolution. In *Companion Volume of the Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 143–146, 2006.

[Preiss, 2001] Judita Preiss. Machine learning for anaphora resolution. 2001.

[Proux *et al.*, 1998] Denys Proux, François Rechenmann, Laurent Julliard, Violaine Pillet, et Bernard Jacq. Detecting gene symbols and names in biological texts : A first step toward pertinent information. In *Proceedings of the Ninth Workshop on Genome Informatics*, pages 72–80, 1998.

[Pyysalo *et al.*, 2006] Sampo Pyysalo, Tapio Salakoski, Sophie Aubin, et Adeline Nazarenko. Lexical adaptation of link grammar to the biomedical sublanguage : a comparative evaluation of three approaches. In Juliane Fluck Sophia Ananiadou, editor, *Proceedings of the Second International Symposium on Semantic Miningin Biomedicine (SMBM 2006)*, pages 60–67, 2006.

[Qiu *et al.*, 2004] Long Qiu, Min-Yen Kan, et Tat-Seng Chua. A public reference implementation of the rap anaphora resolution algorithm. *ArXiv Computer Science e-prints*, pages 291–294, 2004.

[Rabiner, 1989] Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.

[Rich & LuperFoy, 1988] Elaine Rich et Susann LuperFoy. An architecture for anaphora resolution. In *ACL Proceedings, Second Conference on Applied Natural Language Processing*, pages 18–24, 1988.

[Roth & Wen-tau, 2002] Dan Roth et Yih Wen-tau. Probalistic reasoning for entity and relation recognition. In *COLING'02*, 2002.

[Sahami *et al.*, 1998] Mehran Sahami, Susan Dumais, David Heckerman, et Eric Horvitz. A bayesian approach to filtering junk e-mail. *Learning for Text Categorization : Papers from the 1998 Workshop*, 1998.

[Soon *et al.*, 2001] Wee Soon, Hwee Ng, et Daniel Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4) :521–544, 2001.

[sub-committee TC37 SC4, 2005 08 22] ISO sub-committee TC37 SC4. Langage ressource management - morpho-syntactic annotation framework (maf). Technical report, ISO/TC37/SC4/WG2, 2005-08-22.

[Thayse *et al.*, 1990a] André Thayse, Pascal Gribomont, Georges Louis, Dominique Snyers, Pierre Wodon, Paul Gochet, Eric Gregoire, Eduardo Sanchez, et Philippe Delsarte. *Approche logique de l'intelligence artificielle. Tome 1 : De la logique classique à la programmation logique*. Dunod, 1990.

[Thayse *et al.*, 1990b] André Thayse, Pascal Gribomont Guy Hulin, Alain Pirotte, Dominique Roelants, Dominique Snyers, Marc Vauclair, Paul Gochet, Pierre Wolper, Eric Gregoire Georges Louis, et Philippe Delsarte. *Approche logique de l'intelligence artificielle. Tome 2 : de la logique modale à la logique des bases de données*. Dunod, 1990.

[Weissenbacher & Nazarenko, 2007a] Davy Weissenbacher et Adeline Nazarenko. A bayesian approach combining surface clues and linguistic knowledge : Application to the anaphora resolution problem. In *Proceedings of the Recent Advances in Natural Language Processing (RANLP'07)*, 2007.

[Weissenbacher & Nazarenko, 2007b] Davy Weissenbacher et Adeline Nazarenko. Identifier les pronoms anaphoriques et trouver leurs antécédents : l'intérêt de la classification bayésienne. In *Proceedings of Traitement Automatique des Langues Naturelles (TALN'07)*, 2007.

[Witte & Bergler, 2003] René Witte et Sabine Bergler. Fuzzy coreference resolution for summarization. In *Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS)*, pages 43–50, Venice, Italy, June 23–24 2003. Università Ca' Foscari. http://rene-witte.net.

[Zadeh, 1965] Lotfi Zadeh. *Information and Control*, chapter Fuzzy sets, pages 338–353. 1965.